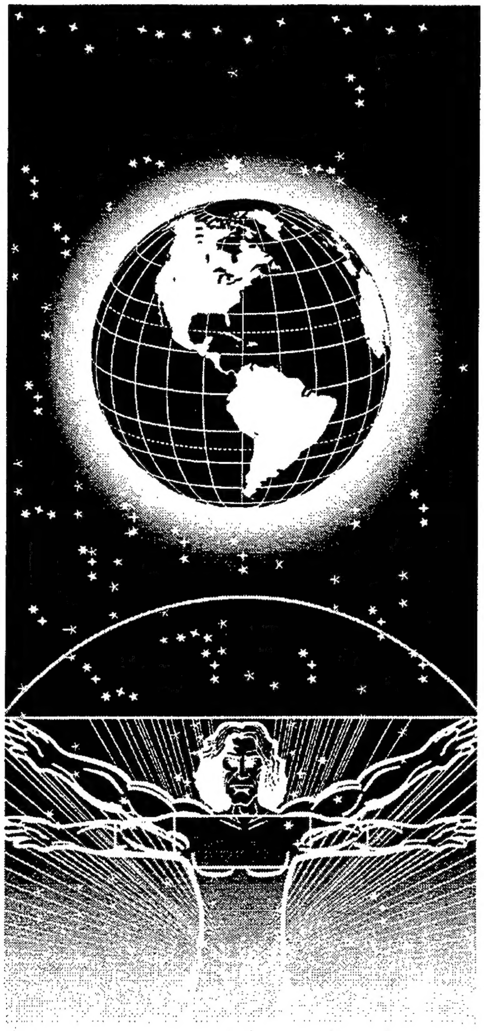
AFRL-HE-BR-TR-1998-0001



AIR FORCE MATERIEL COMMAND
AIR FORCE RESEARCH LABORATORY

REGRESSION TO THE MEAN IN HALF-LIFE STUDIES

Ram C. Tripathi

Division of Mathematics and Computer Science
University of Texas at San Antonio
San Antonio TX 78249

January 1998

Human Effectiveness Directorate
Directed Energy Bioeffects Division
Biomechanisms and Modeling Branch
8111 18th Street
Brooks Air Force Base TX 78235-5215

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this technical report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

Joel E Michalek

JOEL E. MICHALEK, PhD
Contract Monitor

RMiller

RICHARD L. MILLER, PhD
Chief, Directed Energy Bioeffects Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
January 1998

3. REPORT TYPE AND DATES COVERED
Interim - September 1995 - December 1996

4. TITLE AND SUBTITLE

Regression to the Mean in Half-Life Studies

5. FUNDING NUMBERS

C-F41624-96-1-0001
PR-2767
TA-00
WU-F1

6. AUTHOR(S)
Ram C. Tripathi

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Division of Mathematics and Computer Science
University of Texas at San Antonio (UTSA)
San Antonio TX 78285

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Directed Energy Bioeffects Division (AFRL/HED)
Biomechanisms and Modeling Branch (AFRL/HEDB) (AL/AOEP)
2606 Doolittle Road, Bldg 807
Brooks AFB TX 78235-5250

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFRL-HE-BR-TR-1998-0001

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Half-life studies of environmental contaminants in humans are restricted to only a few measurements per subject taken after the initial exposure, the initial dose is usually unknown, and subjects are included in the study only if their body burden is above a threshold c. The assumption of a one compartment first order decay model leads to a repeated measures linear model relating the logarithm of the biomarker with time, with the negative of the coefficient of time being the decay rate. The usual least-squares estimate of the decay rate is biased due to regression to the mean. In this report, based on the repeated measure linear model, unbiased estimates of the decay rate have been developed by the method of least-squares. This has been done for the two cases: (i) when there is no covariate (Report I) and (ii) when there is a categorical covariate (Report II). The maximum likelihood estimator of the decay rate is developed (Report III) under the assumption that the logarithm of the concentration of the contaminant for the k time points of each subject has a truncated multivariate normal distribution with AR (1).

14. SUBJECT TERMS

dioxin, decay rate, estimation
Environmental contaminants

15. NUMBER OF PAGES
48

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UL

CONTENTS

Page

Introduction.. 1

Unbiased Estimation of Regression Parameters Adjusted for Bias Due to Left Truncation...... 3

The Half-Life of Dioxin in Humans Adjusted for a Categorical Covariate.................................17

Maximum Likelihood Estimation for Longitudinal Data with Truncated Observations.............26

References...37

Appendix..39

1 Introduction

Half-life studies of environmental contaminants in humans are generally restricted to only a few measurements taken after the initial exposure. The initial dose is usually unknown because the exposure occurred before the environment was known to be contaminated. We assume that a one compartment first order decay model with decay rate λ holds for subjects with body burden above the background level determined by a threshold c. Subjects are included in the study only if their body burden is greater than c. The threshold is defined to be a high quantile, such as the 99th percentile, of the biomarker distribution in a control population. We assume that the concentration of the contaminant is log-normally distributed which, together with the first order decay model, implies a repeated measures linear model relating the logarithm of the concentration and time, with slope $-\lambda$.

Based on this first order decay model and data for 36 Ranch Hand veterans whose 1982 and 1987 2,3,7,8-tetrachlorodibenzo-p-dioxin (dioxin) measurements were above 10 parts per trillion (ppt), Pirkle et al. (1989) obtained a median dioxin half life of 7.1 years with a 95% confidence interval of 5.8 to 9.6 years. Michalek et al. (1992) used a repeated measures linear model to investigate the effect of the percentage of body fat (PBF) on the decay rate of dioxin. Using weighted least-squares (WLS) estimates, they found a borderline significant association between the decay rate and PBF, with the decay rate of lean subjects being greater than the decay rate of obese subjects. Utilizing the same repeated measures linear model and data collected in 1982, 1987 and 1992, in veterans with more than 10 ppt body burden, Michalek et al. (1996a) obtained an estimate of the decay rate using SAS PROC MIXED and used it to obtain an estimate of the half-life which was corrected for bias. In fact, they obtained the expression for the bias in the estimate of the decay rate utilizing the conditional normal moments from Tallis (1961). They showed that when the data are truncated above a line with a slope $-\lambda$, the WLS estimate of the decay rate becomes unbiased. Their investigation has been carried out under AR(1) and Toeplitz assumptions for the within subject covariance matrices. However, their procedure did not account for covariates such as PBF or age. Thus, it would be of interest to develop an estimate of the decay rate adjusted for covariates. It would also be of interest to develop an estimate of the decay rate using a repeated measures linear model based on truncated data.

In this report we

- develop a WLS estimate of the decay rate based on a repeated measures linear model and correct it for bias assuming a general k, where k represents the number of repeated measures per subject.
- develop a WLS estimate of the decay rate in the above setting in the presence of a categorical covariate.
- develop a maximum likelihood (ML) estimate of the decay rate in the presence of truncation.

In Report I, we present the repeated measures linear model and derive a closed form expression for the WLS estimate of the decay rate. As is well known, WLS estimates are biased if the subjects are selected based on a high or low value of the response variable. We show that by properly adjusting the data, the estimate of λ can be made unbiased. In addition we obtain expressions for other parameter estimates to facilitate a study of their statistical properties.

In Report II, we present the repeated measures linear model which contains a categorical covariate and its interaction with time and derive closed form expressions for the WLS estimates of the associated parameters. Some special cases of the covariance structure are considered. We obtain expressions for the bias in the estimates of the regression parameters under the condition that the body burden of dioxin in the subjects included in the study is greater than a threshold c. The estimates are then made unbiased. The resulting estimates of the half-life are computed using Air Force Health Study (AFHS) data.

In Report III, we derive ML estimates of the parameters under the assumption that the logarithms of dioxin measurements at the k time points for each subject have a truncated multivariate normal distribution. The estimates are obtained by an iterative process similar to the estimation-maximization (EM) algorithm. The WLS estimates serve as initial estimates in the truncated case. An estimate of the asymptotic covariance matrix of the resulting estimators is also derived, which can be used to construct the asymptotic confidence intervals for the parameters.

REPORT I

Unbiased Estimation of Regression Parameters Adjusted for Bias Due to Left Truncation[1]

[1] This research was carried out in collaboration with Dr. Kishan G. Mehrotra, Syracuse University, under the University Resident Research Program of the Air Force Office of Scientific Research.

Abstract

Least-squares estimates of regression parameters are, in general, unbiased. However, if the observations on the response variable are truncated then these estimates become biased due to truncation. For example, half-life studies of environmental contaminants are based on only a few measurements per subject taken after the initial exposure. The subjects are included in the study only if their body burden is above a threshold c. It is assumed that the first order decay model with one compartment holds which leads to a repeated measures linear model relating the logarithm of the contaminant concentration with time and other covariates. The negative of the coefficient of time represents the decay rate (λ), with the half-life of the contaminant given by $t_{1/2} = \ln(2)/\lambda$. The usual WLS estimate of λ is biased due to regression to the mean. We show that the regression parameter can be made unbiased by properly selecting the subjects under study. This selection results in a small loss in the sample size but generally improves the mean-squared error of the estimate.

1 Introduction

Half-life studies of environmental contaminants in humans are generally restricted to only a few measurements taken after the initial exposure. The initial dose is usually unknown because the exposure occurred before the environment was known to be contaminated. We assume that a one compartment first order decay model with decay rate λ holds for subjects with body burden above a background level determined by a threshold c. Subjects are included in the study only if their body burden is greater than c. The threshold is defined to be a high quantile, such as the 99th percentile, of the distribution of the body burden of the contaminant in a control population. We assume that the concentrations are log-normally distributed which, together with the first order decay model, implies a repeated measures linear model relating the logarithm of the concentration and time, with slope $-\lambda$. It is well known that the WLS estimate of λ is biased due to regression to the mean because of the way the subjects were included in the study. It has been shown that (see Michalek et al. (1996a)) if the data are properly conditioned, the WLS estimate of λ can be made unbiased. This process is appealing because unbiased estimates can be obtained through the commercially available packages such as SAS. However, their results were restricted to special cases of the underlying covariance structures and small values of k. In particular, they have shown that if the covariance matrix is AR(1) and $k = 3$, then the samples can be adjusted such that estimate of λ is unbiased. This can be achieved by using SAS without investing in any special purpose computer program.

In this paper we generalize the result of Michalek et al (1996a) for any dimension and show that the samples can be adjusted such that the WLS estimate of λ is unbiased. In addition we obtain expressions for other parameter estimates to facilitate study of their statistical properties.

In section 2, we present the repeated measures linear model and derive closed form expressions for the WLS estimates. Some special cases are considered in section 3. In section 4 we obtain an expression for the bias in the estimate of the regression parameter under the condition that the body burden in the subjects included in the study is greater than a threshold c. The estimates are then made unbiased.

2 Model and Analysis

We assume that k observations were taken per subject for each of n subjects. These subjects were exposed to a contaminant that produced an elevation in the body burden greater than a background level. Suppose that C_0 denotes the initial (unknown) concentration and C_t denotes the concentration t years after the exposure. Then a first-order kinetic model

$$C_t = C_0 e^{-\lambda t} \tag{1}$$

holds in the subjects with body burdens above a threshold c and λ denotes the (unknown) constant decay rate. Based on equation (1), the population half-life is given by

$$t_{1/2} = \frac{\ln(2)}{\lambda}. \tag{2}$$

By taking the natural logarithm of equation (1), we obtain

$$\ln(C_t) = \ln(C_0) - \lambda t. \tag{3}$$

Equation (3) can be regarded as a motivation for a linear regression model with repeated measurements incorporating subject effects.

Let $\boldsymbol{y}_i$ denote the column vector of k observations on the ith subject taken at times $(t_{i1}, t_{i2}, \ldots, t_{ik})$. The regression model, accounting for the subject effects, is given by

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \tau_i + \epsilon_{ij}, \tag{4}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, k$, where $\sum_{i=1}^{n} \tau_i = 0$. Here y_{ij} denotes the natural logarithm of the jth measurement on the ith subject, $-\lambda$ is denoted by β_1, and ϵ_{ij} denotes normal error with mean 0. Our goal is to obtain WLS estimates of β_0, β_1 and the τ_i's.

The inclusion of subjects in the study who have $y_{ij} > \log(c)$ causes left truncation and WLS estimates of the parameters are not necessarily unbiased. WLS estimates of the parameters and the associated bias can be obtained in the vector representation. But, because we want to explore the possibilities of correcting for bias in one or more parameters, it is necessary to obtain their explicit forms. In the following discussion these estimates are obtained and it is shown that the WLS estimate of β_1 can be made unbiased by appropriately selecting the subjects under study.

It is convenient to analyze the problem in terms of vectors of observations for subjects. To that end, we use the following notations.

Notations: For $i = 1, 2, \ldots, n$,

- $\boldsymbol{y}_i$ denotes the k-dimensional vector of observations for the ith subject, $\boldsymbol{y}_i^t = (y_{i1}, y_{i2}, \ldots, y_{ik})$,
- $\boldsymbol{\epsilon}$ denotes the nk-dimensional vector of errors,
- $\boldsymbol{t}_i$ denotes the k-dimensional vector of times for the ith subject, $\boldsymbol{t}_i^t = (t_{i1}, t_{i2}, \ldots, t_{ik})$,
- Φ denotes the covariance matrix of $\boldsymbol{y}_i$,
- $\boldsymbol{Y}$ denotes the column vector of all nk y-observations, $\boldsymbol{Y}^t = [\boldsymbol{y}_1^t, \boldsymbol{y}_2^t, \ldots, \boldsymbol{y}_n^t]$,
- $\boldsymbol{\beta}$ denotes the column vector of all $(n+1)$ parameters, $\boldsymbol{\beta}^t = [\beta_0, \beta_1, \boldsymbol{\tau}^t]$, with $\boldsymbol{\tau}^t = [\tau_1, \tau_2, \ldots, \tau_{n-1}]$,
- $\alpha = \boldsymbol{1}^t \Phi^{-1} \boldsymbol{1}$, where $\boldsymbol{1}$ denotes the $k-$dimensional column vector with all elements equal to 1,
- $t_i^* = \boldsymbol{1}^t \Phi^{-1} \boldsymbol{t}_i$,
- $y_i^* = \boldsymbol{1}^t \Phi^{-1} \boldsymbol{y}_i$,
- $\bar{\boldsymbol{t}} = n^{-1} \sum_{i=1}^n \boldsymbol{t}_i$,
- $\bar{t}^* = n^{-1} \sum_{i=1}^n t_i^*$,
- $S^2 = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{t}_i - \bar{\boldsymbol{t}})^t \Phi^{-1} (\boldsymbol{t}_i - \bar{\boldsymbol{t}})$,
- $CM = \bar{\boldsymbol{t}}^t \Phi^{-1} \bar{\boldsymbol{t}}$.

Thus, the model described in equation (4) can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5}$$

where the design matrix $\boldsymbol{X}$ is

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{t}_1 & \vdots & \boldsymbol{1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{1} & \boldsymbol{t}_2 & \vdots & \boldsymbol{0} & \boldsymbol{1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{1} & \boldsymbol{t}_n & \vdots & -\boldsymbol{1} & -\boldsymbol{1} & \cdots & -\boldsymbol{1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix}.$$

The WLS estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the solution of the system of equations

$$\left(\boldsymbol{X}^t V^{-1} \boldsymbol{X}\right) \widehat{\boldsymbol{\beta}} = \boldsymbol{X}^t V^{-1} \boldsymbol{Y}, \tag{6}$$

where

$$V = \begin{bmatrix} \Phi & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Phi & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Phi \end{bmatrix}.$$

Since $\boldsymbol{X}^t V^{-1} \boldsymbol{X} = \sum_{i=1}^n \boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{X}_i$, and $\boldsymbol{X}^t V^{-1} \boldsymbol{Y} = \sum_{i=1}^n \boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{y}_i$, it follows from (6) that the WLS estimate of $\boldsymbol{\beta}$ satisfies

$$\sum_{i=1}^n \left(\boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{X}_i \right) \widehat{\boldsymbol{\beta}} = \sum_{i=1}^n \boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{y}_i$$

or

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^n \boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{y}_i. \tag{7}$$

Straightforward multiplication gives

$$\sum_{i=1}^n \boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{X}_i = \begin{bmatrix} n\alpha & n\bar{t}^* & \vdots & 0 & 0 & \cdots & 0 \\ n\bar{t}^* & n(S^2 + CM) & \vdots & (t_1^* - t_n^*) & (t_2^* - t_n^*) & \cdots & (t_{n-1}^* - t_n^*) \\ \cdots & \cdots\cdots & \cdots & \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ 0 & (t_1^* - t_n^*) & \vdots & 2\alpha & \alpha & \cdots & \alpha \\ 0 & (t_2^* - t_n^*) & \vdots & \alpha & 2\alpha & \cdots & \alpha \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & (t_{n-1}^* - t_n^*) & \vdots & \alpha & \alpha & \cdots & 2\alpha \end{bmatrix}$$

$$= \begin{bmatrix} A & \vdots & B \\ \cdots & \cdots & \cdots \\ B^t & \vdots & D \end{bmatrix}$$

and

$$\sum_{i=1}^n \boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{y}_i = \begin{bmatrix} \sum_{i=1}^n y_i^* \\ \sum_{i=1}^n \boldsymbol{t}_i^t \Phi^{-1} \boldsymbol{y}_i \\ \cdots \\ y_1^* - y_n^* \\ y_2^* - y_n^* \\ \vdots \\ y_{n-1}^* - y_n^* \end{bmatrix}.$$

The inverse of $\sum_{i=1}^n \boldsymbol{X}_i^t\Phi^{-1}\boldsymbol{X}_i$ can be obtained by using its block representation and expressions for the WLS estimate are found by simplifying the set of linear equations. Let

$$\left(\sum_{i=1}^n \boldsymbol{X}_i^t\Phi^{-1}\boldsymbol{X}_i\right)^{-1} = \begin{bmatrix} P & Q \\ Q & R \end{bmatrix}.$$

Then (see problem 2.8 in Rao (1973)),

$$P = A^{-1} + FE^{-1}F^t, \; Q = -FE^{-1}, \; R = E^{-1},$$

where

$$E = (D - B^tA^{-1}B), \text{ and } F = A^{-1}B.$$

Then,

$$A^{-1} = \frac{n}{\mathcal{D}}\begin{bmatrix} (S^2 + CM) & -\bar{t}^* \\ -\bar{t}^* & \alpha \end{bmatrix},$$

where $\mathcal{D} = n^2[\alpha(S^2 + CM) - (\bar{t}^*)^2]$ is the determinant of A. After substituting for B, D, and A^{-1}, applying problems 2.7 and 2.8 in Rao (1973), and some simplification we get

$$\begin{aligned} R = E^{-1} &= \frac{1}{\alpha}\left[\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right] + \frac{n}{\alpha(\mathcal{D} - n^2s^2)}\mathcal{T}\mathcal{T}^t \\ P &= A^{-1} + \frac{n^3s^2}{\mathcal{D}\alpha(\mathcal{D} - n^2s^2)}\begin{bmatrix} (\bar{t}^*)^2 & -\bar{t}^*\alpha \\ -\bar{t}^*\alpha & \alpha^2 \end{bmatrix} \\ Q &= -\frac{\mathcal{D}}{\alpha\left[\mathcal{D} - n^2s^2\right]}A^{-1}\begin{bmatrix} \mathbf{0}^t \\ \mathcal{T}^t \end{bmatrix}, \end{aligned}$$

where $s^2 = \frac{1}{n}\sum_{i=1}^n (t_i^* - \bar{t}^*)^2$ and $\mathcal{T}^t = \left((t_1^* - \bar{t}^*), (t_2^* - \bar{t}^*), \ldots, (t_{n-1}^* - \bar{t}^*)\right)$.

Expressions for the elements of $\widehat{\boldsymbol{\beta}}$ are obtained by simplifying equation (7). After substituting for the inverse of $\sum_{i=1}^n \boldsymbol{X}_i^t\Phi^{-1}\boldsymbol{X}_i$, we obtain

$$\widehat{\beta_0} = \frac{1}{\alpha}\left[\bar{y}^* - \bar{t}^*\widehat{\beta}_1\right] \tag{8}$$

$$\begin{aligned} \widehat{\beta_1} &= \frac{n}{\left[\mathcal{D} - n^2s^2\right]}\left[\alpha\sum_{i=1}^n \boldsymbol{t}_i^t\Phi^{-1}\boldsymbol{y}_i - \sum_{i=1}^n t_i^*y_i^*\right] \\ &= \frac{n}{\left[\mathcal{D} - n^2s^2\right]}\left[\sum_{i=1}^n \boldsymbol{a}_i^t\boldsymbol{y}_i\right] \end{aligned} \tag{9}$$

and

$$\widehat{\tau} = \begin{bmatrix} y_1^* - \bar{y}^* \\ y_2^* - \bar{y}^* \\ \vdots \\ y_{n-1}^* - \bar{y}^* \end{bmatrix} - \widehat{\beta_1}\begin{bmatrix} t_1^* - \bar{t}^* \\ t_2^* - \bar{t}^* \\ \vdots \\ t_{n-1}^* - \bar{t}^* \end{bmatrix}, \tag{10}$$

where

$$\boldsymbol{a}_i^t = \left[\alpha \boldsymbol{t}_i{}^t - \boldsymbol{1}^t\Phi^{-1}\boldsymbol{t}_i\boldsymbol{1}^t\right]\Phi^{-1} = (a_i[1], a_i[2], \ldots, a_i[k])\,.$$

Since $\boldsymbol{a}_i^t\boldsymbol{1} = \alpha \boldsymbol{t}_i{}^t\Phi^{-1}\boldsymbol{1} - \boldsymbol{1}^t\Phi^{-1}\boldsymbol{t}_i\boldsymbol{1}^t\Phi^{-1}\boldsymbol{1} = 0$, it follows that that $\sum_{j=1}^{k} a_i[j] = 0$ or $\sum_{j=1}^{k-1} a_i[j] = -a_i[k]$. Hence,

$$\widehat{\beta_1} = \frac{n}{[\mathcal{D} - n^2 s^2]}\left[\sum_{i=1}^{n}\sum_{j=1}^{k}\left(a_i[j] - a_i[k]\right)\left(y_{ij} - y_{ik}\right)\right].$$

Because the y_{ij}'s appear in $\widehat{\beta_1}$ only as differences, it will be shown that $\widehat{\beta_1}$ can be made unbiased as long as we can adjust the subjects in the sample. This observation is elaborated in Section 4. First, we mention some special cases of interest.

3 Special Cases

In this section we consider two special cases of the correlation matrix that are the most likely candidates for the repeated measures model. In the first case, we assume that the correlation matrix satisfies the AR(1) assumption and in the second case we relax this assumption and assume that it is given by the Toeplitz matrix. In each case we consider the cases $k = 3$ and 4. We also consider the case when all subjects are exposed to the contaminant at the same time, the rest of the observations are not necessarily at fixed intervals, and the correlation matrix has AR(1) structure.

3.1 AR(1) Correlation Matrix, $k = 3$

Suppose $k = 3$, t_{i1}, t_{i2}, and t_{i3} are equally spaced, with $t_{i2} = t_{i1} + \Delta, t_{i3} = t_{i2} + \Delta$ for all values of i for some fixed Δ, and Φ satisfies AR(1) conditions,

$$\Phi = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}.$$

For convenience, let $t_i \equiv t_{i1}$ for $i = 1, \ldots, n$. It can be seen that in this case

$$nS^2 \;=\; \alpha\sum_{i=1}^{n}(t_i - \overline{t})^2,$$

$$
\begin{aligned}
CM &= \alpha\bar{t}^2 + 2\alpha\bar{t}\,\Delta + \frac{5-4\rho+\rho^2}{1+\rho^2}\Delta^2, \\
ns^2 &= \alpha^2 \sum_1^n (t_i - \bar{t})^2, \\
\bar{t}^* &= \alpha(\bar{t} + \Delta), \\
\alpha &= \frac{3-\rho}{1+\rho},
\end{aligned}
$$

and

$$
\boldsymbol{a}^t = \frac{\alpha\Delta}{(1-\rho^2)}\ [-1,\ \ 0,\ \ 1],
$$

where $\bar{t} = \frac{1}{n}\sum_{i=1}^n t_i$. Consequently,

$$
\begin{aligned}
\mathcal{D} - n^2 s^2 &= \frac{2\alpha}{(1-\rho)^2} n^2 \Delta^2, \\
\widehat{\beta}_1 &= \frac{n(1-\rho)^2(1+\rho)}{2(3-\rho)n^2\Delta^2}\left[\frac{(3-\rho)\Delta}{(1-\rho)^2(1+\rho)} \sum_{i=1}^n (y_{3i} - y_{1i})\right] \\
&= \frac{1}{2n\Delta} \sum_1^n (y_{3i} - y_{1i}), \\
\widehat{\beta}_0 &= \frac{1}{\alpha n} \sum_{i=1}^n [y_{i1} + (1-\rho)y_{i2} + y_{i3}] - (\bar{t} + \Delta)\widehat{\beta}_1,
\end{aligned}
$$

and

$$
\widehat{\tau}_i = \{(y_{i1} + (1-\rho)y_{i2} + y_{i3}) - (y_{i1} + (1-\rho)y_{i2} + y_{i3})\} - \widehat{\beta}_1 \alpha (t_i - \bar{t}).
$$

Michalek et al. (1996a) considered this special case in greater detail.

3.2 AR(1) Correlation Matrix, $k = 4$

We consider the case that $k = 4$ and the t_i's are equally spaced, with $t_{i2} - t_{i1} = t_{i3} - t_{i2} = t_{i4} - t_{i3} = \Delta$. For this reason we denote $t_{i1} \equiv t_i$ and $\bar{t} = n^{-1}\sum_{i=1}^n t_i$. We assume that the covariance structure is given by the AR(1) model

$$
\Phi = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.
$$

It can be seen that

$$
\begin{aligned}
nS^2 &= \alpha \sum_{i=1}^{n} (t_i - \bar{t})^2, \\
\alpha &= \frac{2(2-\rho)}{1+\rho}, \\
CM &= \alpha\bar{t}^2 + \frac{6(2-\rho)}{1+\rho}\bar{t} + \frac{5(1-\rho^2) - 6\rho + 9}{1-\rho^2}\Delta^2, \\
ns^2 &= \alpha^2 \sum_{i=1}^{n} (t_i - \bar{t})^2, \\
t^* &= \alpha\bar{t} + \Delta\frac{3(2-\rho)}{1+\rho},
\end{aligned}
$$

and

$$
\boldsymbol{a}_i^t = \frac{(2-\rho))}{(1+\rho^2)(1-\rho)}\Delta\left[-(3-\rho), \quad -(1-\rho^2), \quad (1-\rho^2), \quad (3-\rho)\right].
$$

Thus,

$$
\widehat{\beta_1} = \frac{1}{(10-5\rho+\rho^2)(n\Delta)} \sum_{i=1}^{n} \left[(3-\rho)(y_{i4} - y_{i1}) + (1-\rho^2)(y_{i3} - y_{i2})\right].
$$

Finally, $\widehat{\beta_0}$ and $\widehat{\tau}$ are obtained from the general forms in equations (8), (9) and (10) with $\bar{t}^*$ defined above and

$$
\begin{aligned}
y_i^* &= \frac{1}{1+\rho}\left[(y_{i1} + y_{i4}) + (1-\rho)(y_{i2} + y_{i3})\right] \\
\bar{y}^* &= \frac{1}{1+\rho}\left[(\bar{y}_1 + \bar{y}_4) + (1-\rho)(\bar{y}_2 + \bar{y}_3)\right].
\end{aligned}
$$

3.3 Toeplitz Correlation Matrix, $k = 3$

As in section 3.1 and 3.2, the t_i's are equally spaced but now we assume that the correlation matrix is Toeplitz of order 3,

$$
\Phi = \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{bmatrix}.
$$

Straightforward simplifications give,

$$
\alpha = \frac{3 - 4\rho_1 + \rho_2}{1 - 2\rho_1 + \rho_2},
$$

$$\begin{aligned}\frac{n}{[\mathcal{D}-n^2s^2]} &= \frac{(1-2\rho_1+\rho_2)(1-\rho_2)}{2n(3-4\rho_1+\rho_2)\Delta^2},\\ -a_i[1] &= a_i[3] = \frac{\Delta(3-4\rho_1+\rho_2)}{(1-2\rho_1+\rho_2)(1-\rho_2)},\\ a_i[2] &= 0.\end{aligned}$$

Consequently, as in the case of AR(1) model,

$$\widehat{\beta_1} = \frac{1}{2n\Delta}\sum_{i=1}^{n}(y_{i3}-y_{i1}).$$

This estimate is equal to the estimate obtained in section 3.1. Estimates of $\widehat{\beta_0}$ and $\widehat{\tau}$ are obtained from the general formulas in (8) and (10).

3.4 Toeplitz Correlation Matrix, $k = 4$

As in the previous case, the t_i's are equally spaced but the correlation matrix has a Toeplitz structure,

$$\Phi = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}.$$

It can be seen that,

$$\begin{aligned}\alpha &= \frac{2-\rho_1-2\rho_2+\rho_3}{1+\rho_1-(\rho_1-\rho_2)^2+\rho_1(\rho_2+\rho_3)},\\ -a_i[1] &= a_i[4] = \frac{(4\rho_1-\rho_2-3)(-2+2\rho_2+\rho_1-\rho_3)}{Q}\Delta,\\ -a_i[2] &= a_i[3] = \frac{(1-3\rho_1+3\rho_2-\rho_3)(2-\rho_1-2\rho_2+\rho_3)}{Q}\Delta,\\ \frac{n}{[\mathcal{D}-n^2s^2]} &= \frac{Q}{n\Delta^2(10-15\rho_1+6\rho_2-\rho_3)(2-\rho_1-2\rho_2+\rho_3)},\end{aligned}$$

where Q denotes the determinant of Φ. Finally,

$$\widehat{\beta_1} = \frac{1}{10-15\rho_1+6\rho_2-\rho_3}\left[(3-4\rho_1+\rho_2)(y_{i4}-y_{i1})+(1-3\rho_1+3\rho_2-\rho_3)(y_{i3}-y_{i2})\right].$$

Other estimates are obtained from the general formulas (8) and (10), and the above estimate of β_1.

3.5 AR(1) Correlation Matrix, $k = 3$, $t_{i1} \equiv t_1$

In this case all t_{i1} are the same, $t_{i1} = t_1$, whereas values of t_{i2} and t_{i3} satisfy $t_1 < t_{i2} < t_{i3}$ and are otherwise arbitrary. We assume that the correlation matrix has AR(1) structure. Then,

$$\begin{aligned}
nS^2 &= \frac{1}{1-\rho^2}\sum_{i=1}^{n}\left[(1+\rho^2)(t_{i2}-\bar{t}_2)^2 - 2\rho(t_{i2}-\bar{t}_2)(t_{i3}-\bar{t}_3) + (t_{i3}-\bar{t}_3)^2\right], \\
CM &= \frac{1}{1-\rho^2}\left[\bar{t}_1^2 - 2\rho\bar{t}_1\bar{t}_2 + (1+\rho^2)\bar{t}_2^2 - 2\rho\bar{t}_2\bar{t}_3 + \bar{t}_3^2\right], \\
ns^2 &= \frac{1}{1+\rho^2}\sum_{i=1}^{n}\left[(1-\rho)(t_{i2}-\bar{t}_2) + (t_{i3}-\bar{t}_3)\right], \\
\bar{t}^* &= \frac{1}{1+\rho}(\bar{t}_1 + \bar{t}_3 + (1-\rho)\bar{t}_2).
\end{aligned}$$

From case 1, we know that $\alpha = (3-\rho)/(1-\rho)$. The vector $\boldsymbol{a}_i$ is

$$\boldsymbol{a}_i = \frac{1}{(1+\rho)(1-\rho^2)}\begin{bmatrix} 2t_1 - (1+\rho)t_{i2} - (1-\rho)t_{i3} \\ (1+\rho)(-t_1 + 2t_{i2} - t_{i3}) \\ -(1-\rho)t_1 - (1+\rho)t_{i2} + 2t_{i3} \end{bmatrix}.$$

Using the fact that $a_i[1] + a_i[2] + a_i[3] = 0$ and

$$\begin{aligned}
\alpha[S^2 + CM] - (\bar{t}^*)^2 - s^2 = \frac{2}{(1+\rho)(1-\rho^2)}&\left[\frac{1}{n}\left((1+\rho)\sum_{i=1}^{n}t_{i2}^2 - (1+\rho)\sum_{i=1}^{n}t_{i2}t_{i3} + \sum_{i=1}^{n}t_{i3}^2\right)\right. \\
&\left. - \left((1+\rho)\bar{t}_1\bar{t}_2 + (1-\rho)\bar{t}_1\bar{t}_3 - \bar{t}_1^2\right)\right],
\end{aligned}$$

we obtain

$$\begin{aligned}
\widehat{\beta_1} = \frac{1}{2n\,[A(t) - B(t)]}\sum_{i=1}^{n}\{&[2t_{i3} - (1+\rho)t_{i2} - (1-\rho)t_{i1}]\,(y_{i3} - y_{i2}) - \\
&[2t_1 - (1+\rho)t_{i2} - (1-\rho)t_{i3}]\,(y_{i2} - y_{i1})\},
\end{aligned}$$

where

$$A(t) = n^{-1}\left\{(1+\rho)\sum_{i=1}^{n}t_{i2}^2 - (1+\rho)\sum_{i=1}^{n}t_{i2}t_{i3} + \sum_{i=1}^{n}t_{i3}^2\right\}$$

and

$$B(t) = \left\{(1+\rho)\bar{t}_1\bar{t}_2 + (1-\rho)\bar{t}_1\bar{t}_3 - \bar{t}_1^2\right\}.$$

Estimates of β_0 and the subject effects are obtained from the general formulas (8) and (10).

4 Bias and bias correction of $\widehat{\beta_1}$

It is well known that if the model (in its matrix formulation) is given by (5), then $\widehat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$. More precisely, if $\mathrm{E}(y_{ij}) = \beta_0 + \beta_1 t_{ij} + \tau_i$, then $\mathrm{E}\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. But, in the current setting, only those y_{ij}'s are included in the sample that satisfy $y_{ij} > \log(c)$. Due to this truncation constraint,

$$\begin{aligned}\mathrm{E}\left(y_{ij} \mid y_{ij} > \log c\right) &= \beta_0 + \beta_1 t_{ij} + \tau_i + \mathrm{E}\left[(y_{ij} - \beta_0 - \beta_1 t_{ij} - \tau_i) \mid y_{ij} > \log c\right] \\ &= \beta_0 + \beta_1 t_{ij} + \tau_i + \mathrm{E}\left[Z_{ij} \mid Z_{ij} > z_{ij}\right],\end{aligned}$$

where Z_{ij} is a normal random variable with mean 0 and $z_{ij} = \log(c) - \beta_0 - \beta_1 t_{ij} - \tau_i$. Hence, $\mathrm{E}\left[y_{ij} \mid y_{ij} > log(c)\right] = \beta_1 + \beta_1 t_{ij} + \tau_{ij}$ only if $z_{ij} = -\infty$. Tallis (1961), and more recently McGill (1992), have evaluated conditional expectations and higher conditional moments of the vector random variable $(Z_{i1}, Z_{i2}, \dots, Z_{ik})$ with correlated components. Thus,

$$\mathrm{E}\left(\widehat{\beta_1}\right) = \beta_1 + \frac{n}{[\mathcal{D} - n^2 s^2]} \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\phi}_i,$$

where $\boldsymbol{\phi}_i = \mathrm{E}\left[(Z_{i1}, Z_{i2}, \dots, Z_{ik})^t \mid Z_{i1} > z_{i1}, Z_{i2} > z_{i2}, \dots, Z_{ik} > z_{ik}\right]$ and $\widehat{\beta_1}$ is a biased estimate of β_1 with bias given by $n/[\mathcal{D} - n^2 s^2] \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\phi}_i$. If we are allowed to *manipulate* the samples, then the bias in $\widehat{\beta_1}$ can be made equal to zero. The basic idea is easy to follow in a special case. Consider the case $k = 3$, equally spaced times, and AR(1) correlation structure. In this case

$$\widehat{\beta_1} = \frac{1}{2n\Delta} \sum_{i=1}^{n} (y_{i3} - y_{i1})$$

and the bias in $\widehat{\beta_1}$ reduces to

$$\mathrm{bias}(\widehat{\beta_1}) = \frac{1}{2n\Delta} \sum_{i=1}^{n} \left(\mathrm{E}(Z_{i3} \mid Z_{i3} > z_{i3}) - \mathrm{E}(Z_{i1} \mid Z_{i1} > z_{i1})\right).$$

This bias can be eliminated provided we can arrange

$$\mathrm{E}(Z_{i3} \mid Z_{i3} > z_{i3}) = \mathrm{E}(Z_{i1} \mid Z_{i1} > z_{i1}),$$

or, equivalently, if we can arrange $z_{i3} = z_{i1}$ for all values of i. If the subjects are included in the sample when $y_{ij} > \log(c)$, then $z_{i3} = \log(c) - \beta_0 - \beta_1 t_i - \tau_i$ and $z_{i1} = \log(c) - \beta_0 - \beta_1 t_i - 2\Delta\beta_1 - \tau_i$ and z_{i3} and z_{i1} are not equal. On the other hand if we are allowed the freedom to shift the truncation points of the y_{i1}'s to $\log(c) - 2\Delta\beta_1$, for $i = 1, \dots, n$, then the new z_{i3} and z_{i1} will be equal and the bias in $\widehat{\beta_1}$ will disappear.

This approach has two drawbacks. First, the (new) truncation point of y_{i1} depends on the unknown parameter β_1. However, this problem can be resolved by repeatedly substituting updated estimates of β_1 until no further change occurs in the value of the estimate of β_1. Second, in this approach subjects are removed from the sample because the new points of truncation are increased. In the AFHS, Michalek et al. (1996b) have found that the reduction in the sample size is small in comparison with the overall size of the sample and the procedure reduces the mean-squared error.

The bias in $\widehat{\beta_1}$ can be removed in the general case in a similar manner. Letting $\boldsymbol{a}_i^t\mathbf{1} = 0$ and $\widehat{\beta_1} = n/[\mathcal{D} - n^2s^2]\sum_{i=1}^n \boldsymbol{a}_i^t\boldsymbol{y}_i$, $\widehat{\beta}_1$ can be written as

$$\widehat{\beta_1} = \frac{n}{[\mathcal{D} - n^2 s^2]}\sum_{i=1}^{n}\sum_{j=2}^{k} [a_i(j) - a_i(k)]\,[y_{ij} - y_{ik}]\,.$$

By arranging the truncation points of the y_{ij}'s such that for all values of i, $z_{ij} = z_{i1}$ for $j = 2, \ldots, k$, the estimate of β_1 will be unbiased. But $z_{ij} = z_{i1}$ can be achieved by shifting the truncation points to $\log(c) + \beta_1(t_{ij} - t_{ik})$ for $j = 2, \ldots, k$, as in Michalek et al. (1996b).

5 Conclusions

In this report we have extended the results of Michalek et al. (1996b) to arbitrary values of k, the number of the repeated observations on each subject. In addition, we have given expressions for the estimates of the other parameters. It remains to be seen if, in the general case, the estimate of β_1 will continue to behave as observed by Michalek et al. (1996a). That is, if the estimate of β_1 is made unbiased, we need to determine whether its mean-squared error be small also, irrespective of the correlation structure.

REPORT II

The Half-Life of Dioxin in Humans Adjusted for a Categorical Covariate[2]

[2] This research was carried out in collaboration with Dr. Kishan G. Mehrotra, Syracuse University, under the University Resident Research Program of the Air Force Office of Scientific Research.

Abstract

Pharmacokinetic studies of environmental contaminants are based on only a few measurements per subject taken after the initial exposure. The subjects are included in the study only if their body burden is above a threshold (c). It is assumed that a first order decay model with one compartment holds, which leads to a repeated measures linear model, relating the logarithm of the concentration of the contaminant with time and other covariates. The negative of the coefficient of time represents the decay rate (λ), with the half-life given by $t_{1/2} = \ln(2)/\lambda$. The usual WLS estimate of the decay rate is biased due to regression to the mean. It has recently been shown (see Michalek et al. (1996b)) that this estimate can be made unbiased if the data are properly conditioned. Since body fat has been found to be an important covariate for predicting the decay rate of dioxin, an unbiased estimate of the decay rate is proposed that is adjusted for an indicator of body fat category and its interaction with time.

1 Introduction

Pharmacokinetic studies of environmental contaminants in humans are generally restricted to only a few measurements taken after the initial exposure. The initial dose is usually unknown, because the exposure occurred before the environment was known to be contaminated. We assume that a one compartment first-order decay model with decay rate λ holds for subjects with body burden above a background level determined by a threshold c. Subjects are included in the study only if their body burden is greater than c. The threshold is defined to be a high quantile, such as 99th percentile, of the distribution of the concentrations of the contaminant in a control population. We assume that the concentrations are log-normally distributed which, together with the first-order decay model, implies a repeated measures linear model relating the logarithm of the biomarker and time, with slope $-\lambda$. It is known that the WLS estimate of λ is biased due to the regression to the mean because of the way the subjects are included in the study. It has been shown (see Michalek et al. (1996b)) that if the data are properly conditioned, the WLS estimate of λ can be made unbiased. This process is appealing, because unbiased estimates can be obtained with commercially available software, such as SAS. However, this estimate is not adjusted for any covariates. For example, because dioxin is lipophilic, body fat is known to be a predictor of the concentration of dioxin. Here we develop a WLS estimate of the decay rate, which is adjusted for binary covariate. This estimate is made unbiased and is used to produce an estimate of the decay rate adjusted for the binary covariate. These results are applied to a pharmacokinetic study of dioxin in veterans of Operation Ranch Hand.

In Section 2 we present the repeated measures linear model and derive closed form expressions for the WLS estimates of the parameters. We also obtain an expression for the bias of the coefficient of time under the condition that the body burden of dioxin in the subjects included in the study is greater than a threshold c. The estimate is then made unbiased. In Section 3, we discuss the estimates in terms of their mean-squared errors. In Section 4, we apply the results to AFHS data.

2 Model and Analysis

Let $\boldsymbol{y}_i$ denote the column vector of k observations on the ith subject. We assume that these observations are taken at k equally spaced times $(t_i, t_i + \Delta, \ldots, t_i + (k-1)\Delta)$. Due to the binary nature of the covariate, the subjects are partitioned into two groups: $g = 1$ and $g = -1$. Without loss of generality we assume that the first n_1 observations belong to group $g = 1$ and the remaining n_2 belong to $g = -1$. Then, a model for the nested design, with subjects nested in groups, is given by

$$y_{ij\ell} = \beta_0 + \beta_1 t_{ij} + \beta_2 g_\ell + \beta_3 g_\ell \times t_{ij} + \tau_{i(\ell)} + \epsilon_{ij(\ell)}, \tag{1}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, k$, $\ell = 1, 2$, where $\tau_{i(\ell)}$ denotes the effect of the ith subject in the ℓth group. We assume that $\sum_{i=1}^{n_\ell} \tau_{i(\ell)} = 0$ for $\ell = 1, 2$. It will be convenient to analyze the problem in terms of vectors of observations for subjects. To that end, we use the following notations.

Notations: For $i = 1, 2, \ldots, n$, and $\ell = 1, 2$,

- $\boldsymbol{y}_i$ denotes the k-dimensional vector of observations for the ith subject,
- $\boldsymbol{t}_i$ denotes the k-dimensional vector of times for the ith subject adjusted for the mean time. In other words,

$$\boldsymbol{t}_i = \begin{pmatrix} t_i \\ t_i + \Delta \\ \vdots \\ t_i + (k-1)\Delta \end{pmatrix} - \begin{pmatrix} \bar{t} + \frac{k-1}{2}\Delta \\ \bar{t} + \frac{k-1}{2}\Delta \\ \vdots \\ \bar{t} + \frac{k-1}{2}\Delta \end{pmatrix} = (t_i - \bar{t})\mathbf{1} + \Delta \boldsymbol{e},$$

where

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \text{ and } \boldsymbol{e} = \frac{1}{2}\begin{pmatrix} -(k-1) \\ -(k-3) \\ \vdots \\ (k-1) \end{pmatrix}$$

are $k-$dimensional vectors, and $\bar{t} = n^{-1}\sum_{i=1}^{n} t_i$.

- We assume that the covariance matrix of $\boldsymbol{y}_i$, denoted by Φ, satisfies AR(1) structure,

$$\Phi = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{k-1} \\ \rho & 1 & \rho & \ldots & \rho^{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \rho^{k-3} & \ldots & 1 \end{bmatrix}.$$

- The vector of all nk y-observations is denoted by $\boldsymbol{Y}$,
$$\boldsymbol{Y}^t = \left[\boldsymbol{y}_1^t, \boldsymbol{y}_2^t, \ldots \boldsymbol{y}_{n_1}^t, \boldsymbol{y}_{n_1+1}^t, \cdots, \boldsymbol{y}_n^t\right].$$
- The vector of all $n+2$ parameters is denoted by $\boldsymbol{\beta}^t = [\beta_0, \beta_1, \beta_2, \beta_3, \boldsymbol{\tau}_1^t, \boldsymbol{\tau}_2^t]$,
- $\alpha = \mathbf{1}^t\Phi^{-1}\mathbf{1}$,
- $\delta = \boldsymbol{e}^t\Phi^{-1}\boldsymbol{e}$,
- $y_i^* = \mathbf{1}^t\Phi^{-1}\boldsymbol{y}_i$,
- $\bar{t_1} = n_1^{-1}\sum_{i=1}^{n_1} t_i$, and $\bar{t_2} = n_2^{-1}\sum_{i=n_1+1}^{n} t_i$, with $n_2 = n - n_1$,
- $\boldsymbol{X}$ denotes the design matrix, and $\boldsymbol{X}_i$ denotes the design matrix associated with the ith subject,
- V denotes the block-diagonal matrix $V = \text{diag}[\Phi, \Phi, \ldots, \Phi]$.

Thus, the model can be written as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{X}$ is given by

$$\boldsymbol{X} = \left[\begin{array}{cccccccccccccc}
\mathbf{1} & t_1 & \mathbf{1} & t_1 & \vdots & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} & \vdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{1} & t_2 & \mathbf{1} & t_2 & \vdots & \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} & \vdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{1} & t_{n_1} & \mathbf{1} & t_{n_1} & \vdots & -\mathbf{1} & -\mathbf{1} & \cdots & -\mathbf{1} & \vdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\mathbf{1} & t_{n_1+1} & -\mathbf{1} & -t_{n_1+1} & \vdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \vdots & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{1} & t_{n_1+2} & -\mathbf{1} & -t_{n_1+2} & \vdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \vdots & \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{1} & t_n & -\mathbf{1} & -t_n & \vdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \vdots & -\mathbf{1} & -\mathbf{1} & \cdots & -\mathbf{1}
\end{array}\right] = \left[\begin{array}{c}
\boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_{n_1} \\ \cdots \\ \boldsymbol{X}_{n_1+1} \\ \boldsymbol{X}_{n_1+2} \\ \vdots \\ \boldsymbol{X}_n
\end{array}\right].$$

The WLS estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, derived from (2), satisfies

$$\left(\boldsymbol{X}^tV^{-1}\boldsymbol{X}\right)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^tV^{-1}\boldsymbol{Y}. \tag{3}$$

Since $\boldsymbol{X}^tV^{-1}\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{X}_i^t\Phi^{-1}\boldsymbol{X}_i$, and $\boldsymbol{X}^tV^{-1}\boldsymbol{Y} = \sum_{i=1}^n \boldsymbol{X}_i^t\Phi^{-1}\boldsymbol{y}_i$, it follows from (3) that $\hat{\boldsymbol{\beta}}$ is given by

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \boldsymbol{X}_i^t\Phi^{-1}\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^n \boldsymbol{X}_i^t\Phi^{-1}\boldsymbol{y}_i. \tag{4}$$

Next, we write

$$\sum_{i=1}^{n} \boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{X}_i = \begin{bmatrix} A & \vdots & B_1 & \vdots & B_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ B_1^t & \vdots & E_1 & \vdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ B_2^t & \vdots & \mathbf{0} & \vdots & E_2 \end{bmatrix} \text{ and } \sum_{i=1}^{n} \boldsymbol{X}_i^t \Phi^{-1} \boldsymbol{y}_i = \begin{bmatrix} W_0 \\ \cdots \\ W_1 \\ \cdots \\ W_2 \end{bmatrix}.$$

Here,

$$A = \begin{bmatrix} A_1 & A_2 \\ A_2 & A_1 \end{bmatrix},$$

for

$$A_1 = \alpha \begin{bmatrix} \sum_{\ell=1}^{2} n_\ell & 0 \\ 0 & \sum_{\ell=1}^{2} n_\ell \, [s_\ell^2 + (\bar{t}_\ell - \bar{t})^2 + \delta\Delta^2] \end{bmatrix},$$

and

$$A_2 = \alpha \begin{bmatrix} \sum_{\ell=1}^{2} (-1)^{\ell+1} n_\ell & \sum_{\ell=1}^{2} (-1)^{\ell+1} \bar{t}_\ell \\ \sum_{\ell=1}^{2} (-1)^{\ell+1} \bar{t}_\ell & \sum_{\ell=1}^{2} (-1)^{\ell+1} n_\ell \, [s_\ell^2 + (\bar{t}_\ell - \bar{t})^2 + \delta\Delta^2] \end{bmatrix},$$

$$B_1 = \alpha \begin{bmatrix} \mathbf{0}^t \\ \mathcal{T}_1^t - (t_{n_1}^* - \bar{t}_1^*) \mathbf{1}_{n_1}^t \\ \mathbf{0}^t \\ \mathcal{T}_1^t - (t_{n_1}^* - \bar{t}_1^*) \mathbf{1}_{n_1}^t \end{bmatrix}, \; B_2 = \alpha \begin{bmatrix} \mathbf{0}^t \\ \mathcal{T}_2^t - (t_n^* - \bar{t}_2^*) \mathbf{1}_{n_2}^t \\ \mathbf{0}^t \\ -\left\{ \mathcal{T}_2^t - (t_n^* - \bar{t}_2^*) \mathbf{1}_{n_2}^t \right\} \end{bmatrix},$$

$$\begin{aligned} W_0 &= \begin{bmatrix} \sum_{i=1}^{n_1} y_i^* + \sum_{i=n_1+1}^{n} y_i^* \\ \sum_{i=1}^{n_1} (t_i - \bar{t}_1) y_i^* + \sum_{i=n_1+1}^{n} (t_i - \bar{t}_2) y_i^* + \Delta \sum_{i=1}^{n} \boldsymbol{e}^t \Phi^{-1} \boldsymbol{y}_i \\ \sum_{i=1}^{n_1} y_i^* - \sum_{i=n_1+1}^{n} y_i^* \\ \sum_{i=1}^{n_1} (t_i - \bar{t}_1) y_i^* - \sum_{i=n_1+1}^{n} (t_i - \bar{t}_2) y_i^* + \Delta \left[\sum_{i=1}^{n_1} \boldsymbol{e}^t \Phi^{-1} \boldsymbol{y}_i - \sum_{i=n_1+1}^{n} \boldsymbol{e}^t \Phi^{-1} \boldsymbol{y}_i \right], \end{bmatrix}, \\ W_1^t &= \left[(y_1^* - y_{n_1}^*), (y_2^* - y_{n_1}^*), \ldots, (y_{n_1-1}^* - y_{n_1}^*) \right], \\ W_2^t &= \left[(y_{n_1+1}^* - y_n^*), (y_{n_1+2}^* - y_n^*), \ldots, (y_{n-1}^* - y_n^*) \right], \end{aligned}$$

where $\mathcal{T}_1^t = [(t_1^* - \bar{t}_1^*), (t_2^* - \bar{t}_1^*), \ldots, (t_{n_1-1}^* - \bar{t}_1^*)]$ and $\mathcal{T}_2^t = [(t_{n_1+1}^* - \bar{t}_2^*), (t_{n_1+2}^* - \bar{t}_2^*), \ldots, (t_{n-1}^* - \bar{t}_2^*)]$. We also define $E_\ell = \alpha \, (\mathbf{I}_\ell + \mathbf{1}_\ell \mathbf{1}_\ell^t)$, where $\mathbf{I}_\ell$ is an identity matrix of size $(n_\ell - 1) \times (n_\ell - 1)$ and $\mathbf{1}_\ell$ is a column vector of size $(n_\ell - 1)$ of all 1's.

We find an explicit expression for $\left(\sum_{i=1}^{n} \mathbf{X}_i^t \Phi^{-1} \mathbf{X}_i\right)^{-1}$. Let

$$\left(\sum_{i=1}^{n} \mathbf{X}_i^t \Phi^{-1} \mathbf{X}_i\right)^{-1} = \begin{bmatrix} A^{-1} + \alpha^2 \mathcal{Q} & \vdots & D_1 & \vdots & D_2 \\ \cdots\cdots\cdots & \cdots & \cdots & \cdots & \cdots \\ D_1^t & \vdots & F_1 & \vdots & 0 \\ \cdots\cdots\cdots & \cdots & \cdots & \cdots & \cdots \\ D_2^t & \vdots & 0 & \vdots & F_2 \end{bmatrix}.$$

Then, using problems 2.7, 2.8, and 2.9 in Rao (1973) repeatedly, we note that A^{-1} and $\mathcal{Q}$ satisfy the following 'block-equality' property,

$$A^{-1} = \begin{bmatrix} a^{11} & a^{12} & \vdots & a^{13} & a^{14} \\ a^{12} & a^{22} & \vdots & a^{23} & a^{24} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a^{13} & a^{14} & \vdots & a^{11} & a^{12} \\ a^{23} & a^{24} & \vdots & a^{12} & a^{22} \end{bmatrix}, \quad \mathcal{Q}^{-1} = \begin{bmatrix} \mathcal{Q}_1 & \vdots & \mathcal{Q}_2 \\ \cdots & \cdots & \cdots \\ \mathcal{Q}_2 & \vdots & \mathcal{Q}_1 \end{bmatrix},$$

where the a^{ij} are elements of A^{-1}, $\mathcal{Q}_1 = n_1 s_1^2 Q_1 + n_2 s_2^2 Q_2$, and $\mathcal{Q}_2 = n_1 s_1^2 Q_1 - n_2 s_2^2 Q_2$. In turn, for $\ell = 1, 2$,

$$Q_\ell = \frac{1}{4n_\ell^2 \alpha \Delta^2 \delta (\alpha s_\ell^2 + \Delta^2 \delta)} \begin{bmatrix} (\bar{t} - \bar{t}_\ell)^2 & (\bar{t} - \bar{t}_\ell) \\ (\bar{t} - \bar{t}_\ell) & 1 \end{bmatrix}.$$

In a similar manner,

$$D_1 = -\alpha \begin{bmatrix} D_{11} \\ D_{11} \end{bmatrix} \mathcal{T}_1 \text{ and } D_2 = -\alpha \begin{bmatrix} D_{12} \\ -D_{12} \end{bmatrix} \mathcal{T}_2,$$

where, for $\ell = 1, 2$,

$$D_{1\ell} = \frac{1}{2\alpha \Delta^2 \delta n_\ell} \begin{bmatrix} (\bar{t} - \bar{t}_\ell) \\ 1 \end{bmatrix}.$$

For $\ell = 1, 2$, let $F_\ell = \frac{1}{\alpha}\left[\mathbf{I}_\ell - \frac{1}{n_\ell}\mathbf{1}_\ell \mathbf{1}_\ell^t\right] + \frac{1}{n_\ell \Delta^2 \delta} \mathcal{T}_\ell \mathcal{T}_\ell^t$.

Substitutions for the values of $A^{-1}, \mathcal{Q}$ and the other terms in equation (4) and simplification gives

$$\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\\ \widehat{\beta}_2\\ \widehat{\beta}_3\end{pmatrix} = \frac{1}{2\alpha}\begin{bmatrix} n_1^{-1}\sum_{i=1}^{n_1} y_i^* + n_2^{-1}\sum_{i=n_1+1}^{n} y_i^* \\ 0 \\ n_1^{-1}\sum_{i=1}^{n_1} y_i^* - n_2^{-1}\sum_{i=n_1+1}^{n} y_i^* \\ 0 \end{bmatrix}$$

$$+\frac{1}{2\Delta\delta}\begin{bmatrix} n_1^{-1}(\bar{t}_1-\bar{t})\sum_{i=1}^{n_1}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i + n_2^{-1}(\bar{t}_2-\bar{t})\sum_{i=n_1+1}^{n}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i \\ n_1^{-1}\sum_{i=1}^{n_1}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i + n_2^{-1}\sum_{i=n_1+1}^{n}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i \\ n_1^{-1}(\bar{t}_1-\bar{t})\sum_{i=1}^{n_1}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i - n_2^{-1}(\bar{t}_2-\bar{t})\sum_{i=n_1+1}^{n}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i \\ n_1^{-1}\sum_{i=1}^{n_1}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i - n_2^{-1}\sum_{i=n_1+1}^{n}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i \end{bmatrix},$$

whereas estimates of the subject effects satisfy

$$\begin{bmatrix}\hat{\boldsymbol{\tau}}_1\\ \hat{\boldsymbol{\tau}}_2\end{bmatrix} = -\frac{1}{2\Delta^2\delta}\begin{bmatrix} n_1^{-1}[(\bar{t}-\bar{t}_1) & 1 & (\bar{t}-\bar{t}_1) & 1]\mathcal{T}_1 \\ n_2^{-1}[(\bar{t}-\bar{t}_2) & 1 & (\bar{t}-\bar{t}_2) & 1]\mathcal{T}_2 \end{bmatrix}\begin{bmatrix} W_{01}+W_{02} \\ W_{01}-W_{02}\end{bmatrix} + \begin{bmatrix} F_1W_1 \\ F_2W_2 \end{bmatrix}.$$

It can be seen that

$$\hat{\boldsymbol{\tau}}_1 = \frac{1}{\alpha}\begin{bmatrix} y_1^*-\bar{y}_1^* \\ y_2^*-\bar{y}_1^* \\ \vdots \\ y_{n_1-1}^*-\bar{y}_1^* \end{bmatrix} - \frac{\sum_{i=1}^{n_1}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i}{n_1\Delta\delta}\begin{bmatrix} t_1-\bar{t}_1 \\ t_2-\bar{t}_1 \\ \vdots \\ t_{n_1-1}-\bar{t}_1 \end{bmatrix},$$

and

$$\hat{\boldsymbol{\tau}}_2 = \frac{1}{\alpha}\begin{bmatrix} y_{n_1+1}^*-\bar{y}_2^* \\ y_{n_1+2}^*-\bar{y}_2^* \\ \vdots \\ y_{n-1}^*-\bar{y}_2^* \end{bmatrix} - \frac{\sum_{i=n_1+1}^{n}\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i}{n_2\Delta\delta}\begin{bmatrix} t_{n_1+1}-\bar{t}_2 \\ t_{n_1+2}-\bar{t}_2 \\ \vdots \\ t_{n-1}-\bar{t}_2 \end{bmatrix}.$$

Special Case

We specialize the above formula for $k=3$. In this case $\alpha=(3-\rho)/(1+\rho), \delta = 2/(1-\rho^2), y_i^* = (y_{i1}+y_{i3}+(1-\rho)y_{i2})/(1-\rho)$ and $\boldsymbol{e}^t\Phi^{-1}\boldsymbol{y}_i = (y_{i3}-y_{i1})/(1-\rho^2)$. Consequently,

$$\beta_1 = \frac{1}{2}\left[\frac{1}{2n_1\Delta}\sum_{i=1}^{n_1}(y_{i3}-y_{i1}) + \frac{1}{2n_2\Delta}\sum_{i=n_1+1}^{n}(y_{i3}-y_{i1})\right]$$

and

$$\beta_3 = \frac{1}{2}\left[\frac{1}{2n_1\Delta}\sum_{i=1}^{n_1}(y_{i3}-y_{i1}) - \frac{1}{2n_2\Delta}\sum_{i=n_1+1}^{n}(y_{i3}-y_{i1})\right].$$

These two estimates are related to the estimates obtained by Michalek et al. (1996b). Their estimate of the regression parameter (based on one sample, without a categorical covariate) is $(2n\Delta)^{-1}\sum(y_{i3}-y_{i1})$. Our estimate can be described as

First, find the estimate of the regression parameter from the first sample (with group value $g = 1$). Second, find the estimate of the regression parameter from the second sample (with group value $g = -1$). Then, a simple average of these two estimates gives an estimate of β_1 and a simple average of the difference of these two estimates gives an estimate of β_3.

The estimates of $\beta_0, \beta_2, \boldsymbol{\tau}_{(1)}$ and $\boldsymbol{\tau}_{(2)}$ also simplify in this case.

3 Unbiasedness and Mean-Squared Errors

The WLS estimates of β_1 and β_3 are biased because the $\boldsymbol{y}_i$'s are left truncated (only those subjects are included in the study for whom the values of the $\boldsymbol{y}_i$'s are greater than $\log(c)$). However, the bias in β_1 and β_3 can be corrected by readjusting the truncation points of the $\boldsymbol{y}_i$'s, as explained in Michalek et al. (1996b). Note that the exercise of fixing the truncation points of the $\boldsymbol{y}_i$'s will be useful only if the mean-squared error of the estimate of β_1 does not increase. Michalek et al. (1996b) have observed that their procedure for correcting the bias actually decreases the mean-squared error.

4 Results for the Air Force Health Study

For the purposes of measuring the change in the decay rate due to PBF category, we divide the data in the AFHS (see Michalek et al. (1996b) and Wolfe et al. (1990)) into two groups. Group 1 consists of subjects with PBF less than the median and Group 2 consists of subjects with PBF greater than the median. The half-life estimates derived from unbiased estimates of the decay rates are

$$\text{Half-life of dioxin in subjects with PBF less than the median} = 7.19 \text{ years}$$
$$\text{Half-life of dioxin in subjects with PBF greater than the median} = 9.72 \text{ years}$$

In contrast, the half-life estimates derived from biased estimates of the decay rate are

$$\text{Half-life of dioxin in subjects with PBF less than the median} = 7.33 \text{ years}$$
$$\text{Half-life of dioxin in subjects with PBF greater than the median} = 10.23 \text{ years}$$

The change in the decay rate due to PBF category is significant. The biased estimates are larger than the unbiased estimates. The change in half-lives due to PBF category is smaller for the unbiased estimates than the biased estimates by approximately a half year.

REPORT III

Maximum Likelihood Estimation for Longitudinal Data with Truncated Observations[3]

[3] This research was carried out in collaboration with Dr. Kishan G. Mehrotra, Syracuse University, under the University Resident Research Program of the Air Force Office of Scientific Research and Dr. Pandu Kulkarni of the University of South Alabama under a grant from The National Research Council.

Abstract

In longitudinal studies, subjects are sometimes included if their measurements at each point of time are above a threshold. However, estimates of parameters of a regression model are often desired in this setting in the presence of covariates. WLS estimates of the regression parameters in the presence of truncation are biased. In this paper, we develop maximum likelihood estimates of the regression parameters when the repeated measurements have a multivariate normal distribution and the data are restricted to lie above a threshold. In addition, the model involves subject effects. The estimates are obtainable by solving a system of nonlinear equations.

1 Introduction

In longitudinal studies, complications may arise by a natural truncation introduced through the subject selection process. For example, Davis (1976) reported a study in which men aged 35 to 39 years were screened for cholesterol levels. Only those men whose cholesterol level exceeded 265 mg% were selected to be in the trial and a treatment was administered to them. Even though it was reasonable to give treatment to only those with a "problem" this selection procedure introduced truncation and the effect of the treatment was confounded with the well known regression to the mean effect. This aspect is well recognized and has been addressed for Davis's data (see Senn and Brown (1985)) and other similar studies (see James (1973)). If a study is well designed then truncation is generally not a problem; for example, consider the study reported by Gardner and Heady (1973). A blood sample was taken at a screening visit and the cholesterol content was analyzed. Subjects participating in the study consisted of the top third of the distribution together with a random half of the bottom third. Subjects in the top third were randomly assigned to the treatment or to a control group (receiving placebo), while the men from the bottom third were given placebo only and formed a second control group. Follow up studies were conducted at 6-month intervals for 2 years and annually afterwards. In this study, the truncation was equally applicable to both the treatment and the placebo groups in the top third group. Therefore, comparison between control and treatment in the top third group was not biased by regression to the mean, whereas a comparison between the treatment group and the bottom third group would have been inappropriate without adjusting for truncation.

The Air Force is conducting a 20-year prospective study of veterans of Operation Ranch Hand, the unit responsible for aerial spraying of Agent Orange and other herbicides in Vietnam from 1962 to 1971. Physical examinations were administered in 1982, 1987 and 1992. Since 1987, exposure has been indexed by a measurement of dioxin in serum. In 1987, all willing Ranch Hand veterans and comparison veterans (who served in Southeast Asia during the same period but who were not involved with spraying) were asked to contribute blood for a dioxin assay; 870 Ranch Hands were assayed and received quantifiable results. As part of a pharmacokinetic study of dioxin, all 343 Ranch Hand veterans with dioxin levels above 10 parts per trillion (ppt) in 1987 and who had stored serum from 1982 examination were selected. Dioxin measurements for these veterans were also made in 1992. In the pharmacokinetic study, only those veterans whose levels were greater than 10 ppt at all three physical examinations were included, resulting in left truncation. The initial dose of a Ranch Hand veteran was unknown because the exposure occurred before the herbicides were known to be contaminated. The goals of the pharmacokinetic study were to find an estimate of the decay rate of dioxin in these veterans and assess the significance of changes in the decay rate with covariates, such as PBF and age.

The AFHS has four special features: (i) the observations are truncated (ii) each subject's observations over the time periods are correlated (iii) a fixed subject effect is necessary and therefore introduces large number of parameters and (iv) there are covariates which may influence the decay rate.

The estimation of the decay rate of dioxin has been the subject of several articles (see Needham et al. (1994), Pirkle et al. (1989), Michalek et al. (1996a), Michalek et al. (1996b)). In all of these studies a one-compartment first order decay model, with decay rate λ, was assumed to hold. If C_{ij} denotes the concentration for subject i t_{ij} years after exposure and C_{i0} is the (unknown) initial concentration, then the first-order kinetics model is given by

$$C_{ij} = C_{i0}\, e^{-\lambda t_{ij}}. \tag{1}$$

Because not all subjects were exposed at the same time, we have considered t_{ij}, the time in years from the initial exposure, to be subject-dependent. If we take the natural logarithm of (1), we obtain

$$\ln(C_{ij}) = \ln(C_{i0}) - \lambda t_{ij}. \tag{2}$$

Equation (2) motivates a repeated measures linear model on the log scale, with $\beta_1 = -\lambda$,

$$y_{ij} = \beta_0 + \tau_i + \beta_1 t_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots, k. \tag{3}$$

A WLS procedure is an obvious choice for estimation, however, WLS estimates will be biased if the observations are truncated.

The magnitude of the bias can be studied by means of a closed form expression for the WLS estimate of β_1. Michalek et al. (1996a) have described an ad hoc procedure that iteratively results in an approximately unbiased estimator. Michalek et al. (1996b) derived a closed form estimate and used it to estimate the bias. But, since their formula for bias also contained the parameter they were trying to estimate, it was difficult to assess the accuracy of the estimate of the bias. Another approach is to consider a procedure that accommodates the truncation and provides estimators that are relatively easy to compute, are at least asymptotically unbiased, and are asymptotically normally distributed. In this paper, we give an alternative method to provide closed form expressions for the WLS estimates of the parameters of a regression model with repeated measures and fixed effects when the observations are truncated. We also obtain ML estimates of the parameters based on the truncated observations. The asymptotic properties of the estimators follow from standard maximum likelihood theory. We consider two procedures to estimate the variances of the ML estimates, one via the bootstrap method and the other via the inversion of the information matrix.

In section 2, we introduce the model in general form and obtain the estimating equations for the case of truncated observations. We obtain closed form expressions for the WLS estimates for the untruncated case and provide ML estimates for the case of truncated observations. Section 3, we derive the variance-covariance matrix of the estimates via inversion of the information matrix. We apply the results to AFHS data in Section 4.

2 Maximum Likelihood Estimation

We consider estimation of the parameters of the regression model with fixed subject effects and with repeated measures in the context of left truncation. A similar development can be addressed for right or middle truncation. Consider model (3) in vector notations. Let $\boldsymbol{y}_i$ be the random variable representing the k observations associated with the ith subject. Then, model (3) can be written as

$$\boldsymbol{y}_i = \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n,$$

where we assume that the errors $\boldsymbol{\epsilon}_i$ follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Ω, and $\boldsymbol{\theta}_i$ is given by

$$\boldsymbol{\theta}_i = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{t}_i + \tau_i \mathbf{1}, \tag{4}$$

where $\boldsymbol{t}_i^t = (t_{i1}, t_{i1}, \ldots, t_{ik})$ and $\mathbf{1}^t = (1, \ldots, 1)$ are k-dimensional column vectors and the τ_i's satisfy the restriction $\tau_1 + \tau_2 + \ldots + \tau_n = 0$.

The ML estimates are derived under the assumption that $\boldsymbol{y}_i > \boldsymbol{a}$ for all values of i, where $\boldsymbol{a}^t = (a, a, \ldots, a)$. Regression models with truncation have been discussed extensively in the literature; see Maddala (1983). But to the best of our knowledge, a general discussion of the repeated measures model with fixed effects and truncation has not received much attention. Let $S(\boldsymbol{a}, \boldsymbol{\theta}_i)$ be the survival function for the ith subject. Then, the conditional likelihood function and the corresponding log-likelihood function are given by

$$L = \prod_{i=1}^{n} \left[(2\pi)^{-\frac{k}{2}} \mid \Omega \mid^{-\frac{1}{2}} \exp[-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\theta}_i)^t \, \Omega^{-1} \, (\boldsymbol{y}_i - \boldsymbol{\theta}_i)] / S(\boldsymbol{a}, \boldsymbol{\theta}_i) \right],$$

$$\equiv \prod_{i=1}^{n} f(\boldsymbol{y}_i, \boldsymbol{\theta}_i)/S(\boldsymbol{a}, \boldsymbol{\theta}_i),$$

and

$$\ln(L) = \sum_{i=1}^{n} \ln(L_i) = \sum_{i=1}^{n} \ln(f(\boldsymbol{y}_i, \boldsymbol{\theta}_i)) - \sum_{i=1}^{n} \ln(S(\boldsymbol{a}, \boldsymbol{\theta}_i)), \tag{5}$$

where

$$S(\boldsymbol{a}, \boldsymbol{\theta}_i) = \int_{\boldsymbol{a}}^{\infty} f(\boldsymbol{y}_i, \boldsymbol{\theta}_i) \mathrm{d}\boldsymbol{y}_i.$$

Equation (5) has two components, the first represents the 'usual' log-likelihood without the truncation effect, whereas the second component is due to truncation. Under the assumption of normality, the ML estimates obtained by using the first component are the same as the WLS estimates. To simplify the presentation, we first derive the WLS estimates.

2.1 The weighted least-squares estimates

The WLS estimates of the parameters are obtained by minimizing

$$Q = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta}_i)^t \Omega^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}_i).$$

By differentiating Q with respect to β_0, β_1 and the τ_i's, using the chain rule and noting from (4) that

$$\frac{\partial \boldsymbol{\theta}_i}{\partial \beta_0} = \mathbf{1}, \quad \frac{\partial \boldsymbol{\theta}_i}{\partial \beta_1} = \boldsymbol{t}_i, \quad \frac{\partial \boldsymbol{\theta}_i}{\partial \tau_i} = \mathbf{1}, \tag{6}$$

and equating the derivatives to zero, we find, using (6), that

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= \sum_{i=1}^{n} \mathbf{1}^t \Omega^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}_i) = 0, \\ \frac{\partial Q}{\partial \beta_1} &= \sum_{i=1}^{n} \boldsymbol{t}_i^t \Omega^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}_i) = 0, \end{aligned}$$

and, for $i = 1, \ldots, n-1$,

$$\frac{\partial Q}{\partial \tau_i} = \mathbf{1}^t \Omega^{-1} [(\boldsymbol{y}_i - \boldsymbol{y}_n) - (\boldsymbol{\theta}_i - \boldsymbol{\theta}_n)] = 0.$$

Substitution for $\boldsymbol{\theta}_i$ in terms of β_0, β_1 and the τ_i's and a straightforward simplification gives

$$\begin{aligned}
\sum_{i=1}^{n} \mathbf{1}^t\Omega^{-1}\boldsymbol{y}_i &= n(\mathbf{1}^t\Omega^{-1}\mathbf{1})\beta_0 + (\sum_{i=1}^{n} \mathbf{1}^t\Omega^{-1}\boldsymbol{t}_i)\beta_1, \\
\sum_{i=1}^{n} \boldsymbol{t}_i^t\Omega^{-1}\boldsymbol{y}_i &= (\sum_{i=1}^{n} \mathbf{1}^t\Omega^{-1}\boldsymbol{t}_i)\beta_0 + (\sum_{i=1}^{n} \boldsymbol{t}_i^t\Omega^{-1}\boldsymbol{t}_i)\beta_1 + \sum_{i=1}^{n}(\mathbf{1}^t\Omega^{-1}\boldsymbol{t}_i)\tau_i,
\end{aligned}$$

and, for $i = 1, \ldots, n-1$,

$$\mathbf{1}^t\Omega^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_n) = \mathbf{1}^t\Omega^{-1}(\boldsymbol{t}_i - \boldsymbol{t}_n)\beta_1 + \mathbf{1}^t\Omega^{-1}\mathbf{1}(\tau_i - \tau_n).$$

The key to obtaining a solution for the above system of equations is to first obtain an expression for τ_i, $i = 1, \ldots, n-1$, in terms of the other parameters, substitute it in the first two equations, and solve for β_0 and β_1. The WLS estimates of β_0, β_1 and τ_i for $i = 1, \ldots, n-1$ are

$$\begin{aligned}
\widehat{\beta_0} &= \frac{1}{n\alpha}\left[\sum_{i=1}^{n} \mathbf{1}^t\Omega^{-1}\boldsymbol{y}_i\right] - \frac{1}{n\alpha}\sum_{i=1}^{n}(\mathbf{1}^t\Omega^{-1}\boldsymbol{t}_i)\widehat{\beta_1} \\
\widehat{\beta_1} &= \frac{\alpha\sum_{i=1}^{n} \boldsymbol{t}_i^t\Omega^{-1}\boldsymbol{y}_i - \sum_{i=1}^{n}(\mathbf{1}^t\Omega^{-1}\boldsymbol{y}_i)(\mathbf{1}^t\Omega^{-1}\boldsymbol{t}_i)}{\alpha\sum_{i=1}^{n} \boldsymbol{t}_i^t\Omega^{-1}\boldsymbol{t}_i - \sum_{i=1}^{n}(\mathbf{1}^t\Omega^{-1}\boldsymbol{t}_i)(\mathbf{1}^t\Omega^{-1}\boldsymbol{t}_i)} \qquad (7) \\
\widehat{\tau_i} &= \frac{1}{\alpha}\left[\mathbf{1}^t\Omega^{-1}(\boldsymbol{y}_i - \overline{\boldsymbol{y}})\right] - \frac{1}{\alpha}\left[\mathbf{1}^t\Omega^{-1}(\boldsymbol{t}_i - \overline{\boldsymbol{t}})\right]\widehat{\beta_1},
\end{aligned}$$

where $\alpha = \mathbf{1}^t\Omega^{-1}\mathbf{1}$.

Remark 1: In the absence of fixed subject effects, the τ_i's, these normal equations are similar to the "classical" normal equations of linear regression. The main difference is that all k observations associated with the ith subject are represented by a scalar variable $u_i = \mathbf{1}^t\Omega^{-1}\boldsymbol{y}_i$. Similar scalar reductions are obtained for the $\boldsymbol{t}_i$'s. Thus repeated measures regression can be treated essentially as ordinary simple regression with a change of variables.

Remark 2: The case that $\boldsymbol{t}_i = t_i\mathbf{1} + \Delta\boldsymbol{e}$ represents successive observations taken at equally spaced intervals of length Δ for each subject. Let $\boldsymbol{e}^t = (0, 1, \ldots, (k-1))$. In this case, (7) simplifies to

$$\widehat{\beta_1} = \frac{\alpha\sum_{i=1}^{n} \boldsymbol{e}^t\Omega^{-1}\boldsymbol{y}_i - (\mathbf{1}^t\Omega^{-1}\boldsymbol{e})\sum_{i=1}^{n}(\mathbf{1}^t\Omega^{-1}\boldsymbol{y}_i)}{n\alpha\Delta(\boldsymbol{e}^t\Omega^{-1}\boldsymbol{e} - \alpha)}.$$

Remark 3: For $k = 3$, if Ω has AR(1) structure, then (7) simplifies to

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(y_{3i} - y_{1i})}{2n\Delta}.$$

This special case was first obtained by Michalek et al.(1996a). Further, they obtained the same expression when Ω is a Toeplitz matrix of order 3. The expression for $\widehat{\beta_1}$ for larger values of k when Ω is either an AR(1) or Toeplitz involves elements of Ω.

Remark 4 The WLS estimate of β_1 is biased because truncated values of $\boldsymbol{y}_i$ are used in the estimate. In general, the bias is

$$\text{bias}(\widehat{\beta_1}) = \frac{\alpha \sum_{i=1}^{n} \boldsymbol{t}_i^t \Omega^{-1} \boldsymbol{H}_i - \sum_{i=1}^{n} (\mathbf{1}^t \Omega^{-1} \boldsymbol{H}_i (\mathbf{1}^t \Omega^{-1} \boldsymbol{t}_i)}{\alpha \sum_{i=1}^{n} \boldsymbol{t}_i^t \Omega^{-1} \boldsymbol{t}_i - \sum_{i=1}^{n} (\mathbf{1}^t \Omega^{-1} \boldsymbol{t}_i)(\mathbf{1}^t \Omega^{-1} \boldsymbol{t}_i)},$$

where $\boldsymbol{H}_i \equiv \boldsymbol{H}_i(\boldsymbol{a} - \boldsymbol{\theta}_i)$ is the multivariate hazard vector defined in McGill [(1992), equation (9)]. This bias was obtained by Michalek et al. (1996b) for the special case when $k = 3$ and Ω has an AR(1) structure. However, the formula for the bias depends on β_1. In the absence of an unbiased estimate of β_1 it is difficult to assess the accuracy of the estimate of this bias. ML estimation, discussed below, adjusts for truncation through the second expression of equation (8).

2.2 Maximum likelihood estimates

The WLS estimates were obtained by ignoring the second component of the likelihood expression in (5). The ML estimates on the other hand use the entire likelihood, taking into account the truncation component as well. It can be verified that

$$\frac{\partial \ln(L_i)}{\partial \boldsymbol{\theta}_i} = \Omega^{-1}(\boldsymbol{y}_i - \boldsymbol{\theta}_i) - \Omega^{-1}\mathrm{E}(\boldsymbol{y}_i - \boldsymbol{\theta}_i \mid \boldsymbol{y}_i > \boldsymbol{a}),$$

where $\mathrm{E}(\boldsymbol{y}_i - \boldsymbol{\theta}_i \mid \boldsymbol{y}_i > \boldsymbol{a})$ denotes the conditional expectation of $\boldsymbol{y}_i - \boldsymbol{\theta}_i$ given $\boldsymbol{y}_i > \boldsymbol{a}$. Consequently, the ML estimates of β_0, β_1 and the τ_i's are solutions of the system of equations

$$\begin{aligned}
\frac{\partial \ln L}{\partial \beta_0} &= \sum_{i=1}^{n} \mathbf{1}^t \Omega^{-1}(\boldsymbol{y}_i - \boldsymbol{\theta}_i) - \sum_{i=1}^{n} \mathbf{1}^t \Omega^{-1}\mathrm{E}(\boldsymbol{y}_i - \boldsymbol{\theta}_i \mid \boldsymbol{y}_i > \boldsymbol{a}) = 0, \\
\frac{\partial \ln L}{\partial \beta_1} &= \sum_{i=1}^{n} \boldsymbol{t}_i^t \Omega^{-1}(\boldsymbol{y}_i - \boldsymbol{\theta}_i) - \sum_{i=1}^{n} \boldsymbol{t}_i^t \Omega^{-1}\mathrm{E}(\boldsymbol{y}_i - \boldsymbol{\theta}_i \mid \boldsymbol{y}_i > \boldsymbol{a}) = 0,
\end{aligned} \tag{8}$$

and, for $i = 1, \ldots, n-1$,

$$\begin{aligned}
\frac{\partial \ln L}{\partial \tau_i} &= \mathbf{1}^t \Omega^{-1}[(\boldsymbol{y}_i - \boldsymbol{y}_n) - (\boldsymbol{\theta}_i - \boldsymbol{\theta}_n)] \\
&\quad - \mathbf{1}^t \Omega^{-1}[\mathrm{E}(\boldsymbol{y}_i - \boldsymbol{\theta}_i \mid \boldsymbol{y}_i > \boldsymbol{a}) - \mathrm{E}(\boldsymbol{y}_n - \boldsymbol{\theta}_n \mid \boldsymbol{y}_n > \boldsymbol{a})] = 0.
\end{aligned}$$

The second components of (8) introduce non-linearity. A solution is obtained by an iterative procedure, described below.

Step 1. Use WLS estimates of the parameters as the initial estimates $\widehat{\beta_0}(0)$, $\widehat{\beta_1}(0)$ and $\widehat{\tau}_i(0)$, $i = 1, \ldots, n-1$, and obtain $\widehat{\boldsymbol{\theta}_i}(0)$.

Step 2.

(a) Evaluate[4] $\mathrm{E}\left[\boldsymbol{y}_i - \widehat{\boldsymbol{\theta}_i}(0) \mid \boldsymbol{y}_i > \boldsymbol{a}\right]$ for each value of i.

(b) Substitute these expected values in (8) to get a new set of equations which are similar to the equations for the WLS estimates with $\boldsymbol{y}_i$ replaced by

$$\boldsymbol{y}_i^* = \boldsymbol{y}_i + \mathrm{E}\left[\boldsymbol{y}_i - \widehat{\boldsymbol{\theta}_i}(0) \mid \boldsymbol{y}_i > \boldsymbol{a}\right]$$

for each value of i. Using $\boldsymbol{y}_i^*$ in place of $\boldsymbol{y}_i$ in (7) obtain $\widehat{\beta_0}(1), \widehat{\beta_1}(1)$ and $\widehat{\tau}_i(1)$ for $i = 1, \ldots, n-1$, and $\widehat{\boldsymbol{\theta}_i}(1)$.

Step 3. Repeat Step 2 with $\widehat{\boldsymbol{\theta}_i}(0)$ replaced by $\widehat{\boldsymbol{\theta}_i}(1)$. Repeat this procedure until the difference between two consecutive estimates is negligible.

Our ML procedure is similar to the EM algorithm. Navidi (1997) gives a simple graphical illustration of the EM algorithm. In the classical EM algorithm, the expectation of the unobservable component of the log likelihood function is obtained, whereas we estimate the conditional expected value by the most recent estimate of the unknown parameter $\boldsymbol{\theta}_i$.

3 Estimates of Information Matrix and the Variance of ML Estimates

The ML estimates obtained in the previous section fall into the standard maximum likelihood class. Properties of the estimates follow from standard results. The asymptotic distribution of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ is bivariate normal with covariance given by the inverse of the Fisher information matrix. The information matrix is estimated by evaluating the negative of the second derivative of the log-likelihood at the ML estimates, denoted by $\mathcal{I}$. For our model this matrix is an $(n+1) \times (n+1)$ matrix. To obtain the elements of this matrix we use the chain rule of differentiation and equation (6). To this end, we find that

$$-\frac{\partial^2 \ln L_i}{\partial \boldsymbol{\theta}_i \boldsymbol{\theta}_i^t} = \Omega^{-1} \Psi(\boldsymbol{a}, \boldsymbol{\theta}_i) \Omega^{-1},$$

[4] A method to evaluate these and similar conditional expectations is given in the Appendix.

where $\Psi(\boldsymbol{a},\boldsymbol{\theta}_i)$ denotes the conditional covariance matrix of $\boldsymbol{y}_i - \boldsymbol{\theta}_i$, given $\boldsymbol{y}_i > \boldsymbol{a}$. That is,

$$\Psi(\boldsymbol{a},\boldsymbol{\theta}_i) = \mathrm{E}\left[(\boldsymbol{y}_i - \boldsymbol{\theta}_i)(\boldsymbol{y}_i - \boldsymbol{\theta}_i)^t \mid \boldsymbol{y}_i > \boldsymbol{a}\right] - \mathrm{E}\left[(\boldsymbol{y}_i - \boldsymbol{\theta}_i) \mid \boldsymbol{y}_i > \boldsymbol{a}\right] \mathrm{E}\left[(\boldsymbol{y}_i - \boldsymbol{\theta}_i)^t \mid \boldsymbol{y}_i > \boldsymbol{a}\right].$$

Application of (6) gives

$$\begin{aligned}
-\frac{\partial^2 \ln L}{\partial \beta_0^2} &= \sum_{i=1}^{n} \mathbf{1}^t \Omega^{-1} \Psi(\boldsymbol{a},\boldsymbol{\theta}_i) \Omega^{-1} \mathbf{1} \\
-\frac{\partial^2 \ln L}{\partial \beta_1^2} &= \sum_{i=1}^{n} \boldsymbol{t}_i^t \Omega^{-1} \Psi(\boldsymbol{a},\boldsymbol{\theta}_i) \Omega^{-1} \boldsymbol{t}_i \\
-\frac{\partial^2 \ln L}{\partial \beta_0 \partial \beta_1} &= \sum_{i=1}^{n} \mathbf{1}^t \Omega^{-1} \Psi(\boldsymbol{a},\boldsymbol{\theta}_i) \Omega^{-1} \boldsymbol{t}_i.
\end{aligned}$$

Likewise, for $i = 1, \ldots, n-1,\ j = 1 \ldots, n-1,$

$$\begin{aligned}
-\frac{\partial^2 \ln L}{\partial \tau_i^2} &= \mathbf{1}^t \Omega^{-1} \left\{\Psi(\boldsymbol{a},\boldsymbol{\theta}_i) + \Psi(\boldsymbol{a},\boldsymbol{\theta}_n)\right\} \Omega^{-1} \mathbf{1} \\
-\frac{\partial^2 \ln L}{\partial \tau_i \partial \tau_j} &= \mathbf{1}^t \Omega^{-1} \Psi(\boldsymbol{a},\boldsymbol{\theta}_n) \Omega^{-1} \mathbf{1} \\
-\frac{\partial^2 \ln L}{\partial \beta_0 \partial \tau_i} &= \mathbf{1}^t \Omega^{-1} \left\{\Psi(\boldsymbol{a},\boldsymbol{\theta}_i) - \Psi(\boldsymbol{a},\boldsymbol{\theta}_n)\right\} \Omega^{-1} \mathbf{1} \\
-\frac{\partial^2 \ln L}{\partial \beta_1 \partial \tau_i} &= \mathbf{1}^t \Omega^{-1} \Psi(\boldsymbol{a},\boldsymbol{\theta}_i) \Omega^{-1} \boldsymbol{t}_i - \mathbf{1}^t \Omega^{-1} \Psi(\boldsymbol{a},\boldsymbol{\theta}_n) \Omega^{-1} \boldsymbol{t}_n.
\end{aligned}$$

Expressions containing conditional moments can be evaluated by numerical integration. The case that $k = 3$ and Ω has an AR(1) structure has been described in detail in the Appendix. Due to the special nature of the information matrix, shown below, its inversion can be easily computed. We write $\mathcal{I}$ in block form as

$$\mathcal{I} = \begin{bmatrix} \mathcal{I}_{11} & \vdots & \mathcal{I}_{12} \\ \cdots & \cdots & \cdots \\ \mathcal{I}_{21} & \vdots & \mathcal{I}_{22} \end{bmatrix},$$

where $\mathcal{I}_{11}$ represents the information expression associated with β_0 and β_1, $\mathcal{I}_{22}$ represents the information expression associated with τ_i, for $i = 1, \ldots, n-1$, and $\mathcal{I}_{12} = \mathcal{I}_{21}^t$ represents the cross-information expressions. Then,

$$\mathcal{I}_{22} = \mathrm{diag}[\varrho_1, \ldots, \varrho_{n-1}] + \varrho_n \mathbf{1}_{n-1} \mathbf{1}_{n-1}^t,$$

where $\varrho_i = \mathbf{1}^t \Omega^{-1} \Psi(\boldsymbol{a},\boldsymbol{\theta}_i) \Omega^{-1} \mathbf{1}$ and $\mathbf{1}_{n-1}$ denotes an $(n-1)$-dimensional vector of all 1's. Using this special property of $\mathcal{I}_{22}$ and the block inverse approach described in problems 2.7 and 2.8 in Rao (1973), the asymptotic variances and covariances can be estimated for all parameters. In other words, the numerical inversion of $\mathcal{I}$, whose size is $(n+1) \times (n+1)$, is not necessary; an estimate of the asymptotic covariance matrix only requires the inversion of a 2×2 matrix.

The Bootstrap Method:

An alternative to the matrix inversion method of covariance estimation is the bootstrap. The matrix inversion method depends second order conditional moments, whereas the bootstrap method is computationally intensive. The bootstrap method will be studied in a future report.

4 Discussion

In order to obtain the variances of the intercept and slope estimators we needed to invert the information matrix (which in our case was a 241×241 matrix). However, we made use of the special structures which reduced the inversion problem into inverting a 2×2 matrix. A comparison of the ML estimate of dioxin half-life and it's standard deviation with the WLS estimate and standard deviation indicates that the ML estimate of the half-life is about 1 year shorter (7.3 years) and the standard deviation of the ML estimate is slightly larger than that of the WLS method. Although we have concentrated on a special covariance structure, our results are general and can be used for other covariance structures as well. The complexity of the problem is essentially in the numerical evaluation of the conditional expectation needed in Step (2) of the ML procedure. The regression model can be extended to include other covariates. If there is no truncation then standard software will provide the WLS or ML estimates and their standard errors. If the observations are truncated, the ML method will require special purpose programs. In the context of truncated observations the WLS estimates are generally biased and the maximum likelihood provides more reliable estimates, at least when the sample size is large. If the sample size is small the ML estimates can be unreliable, so one may resort to procedures such as the bootstrap and the jackknife, whereas the WLS estimates are unbiased regardless of the sample size.

Even though the number of equations to be solved is minimal, because of the truncation, the equations are non-linear in the parameters. Therefore they must be solved iteratively.

References

[1] Davis, C. E. (1976). The effect of regression to the mean in epidemiologic and clinical studies. *American Journal of Epidemiology* **104**, 493–498.

[2] Gardner, M. J. and Heady, J. A. (1973). Some effects of within-person variability in epidemiological studies. *Journal of Chronic Diseases* **26**, 781–795.

[3] James, K. E. (1973). Regression toward the mean in uncontrolled clinical studies. *Biometrics* **29**, 121–130.

[4] Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics.* Cambridge University Press, Cambridge,

[5] McGill, J. I. (1992). The multivariate hazard gradient and moments of the truncated multinormal distribution. *Communications in Statistics, Theory and Methods* **21**, 3053–3060.

[6] Michalek, J. E., Tripathi, R. C., Caudill, S. P., and Pirkle, J. L. (1992). Investigation of dioxin half-life heterogeneity in veterans of Operation Ranch Hand. *J. Toxicology and Environmental Health* **35**, 29-38.

[7] Michalek, J. E., Pirkle, J. L., Caudill, S. P., Tripathi, R. C., Patterson, D. G., Jr., and Needham, L. L. (1996a). Pharmacokinetics of of dioxin in veterans of Operation Ranch Hand; 10 year follow-up. *J. Toxicology and Environmental Health* **47** 209-220.

[8] Michalek, J. M., Tripathi, R. C., Kulkarni, P., Selvavel, K. and Gupta, P. L. (1996b). Correction to bias introduced by truncation in pharmacokinetic studies of environmental contaminants. (*Environmetrics*, to appear).

[9] Navidi, W. (1997). A graphical illustration of the EM algorithm. *The American Statistician* **51**, 29–31.

[10] Needham, L. L., Gerthoux, P. M., Patterson, D. G., Jr., Brambilla, P., Pirkle, J. L., Tramacere, P. I., Turner, W. E., Beretta, C., Sampson, E. J., and Mocarelli, P. (1994). Half-life of 2,3,7,8-tetrachlorodibenzo-p-dioxin in serum of Seveso adults: Interim Reports. *Organolhalogen Compounds* **21** 81-85.

[11] Pirkle, J. L., Wolfe, W. H. Patterson, D. G., Jr., Needham, L. L., Michalek, J. E., Miner, J. C., and Peterson, M. R. (1989). Estimates of the half-life of 2,3,7,8-tetrachlorodibenzo-p-dioxin in Vietnam veterans of Operation Ranch Hand. *J. Toxicology and Environmental Health* **27** 165-171.

[12] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, **2nd** edition, Wiley & Sons, New York.

[13] Senn, S. J. and Brown, R. A. (1985). Estimating treatment effects in clinical trials subject to regression to mean. *Biometrics* **41**, 555–560.

[14] Shepard, D. S. and Finison, L. J. (1983). Blood pressure reductions: correcting for regression to the mean. *Preventive Medicine* **12**, 304–317.

[15] Tallis, G. M. (1961). The moment generating function of the truncated multinormal distribution. *Journal of the Royal Statistical Society* **B**, **23**, 223–229.

[16] Wolfe, W. H., Michalek, J. E., Miner, J. C., Rahe, A. J., Silva, J., Thomas, W. F., Grubbs, W. D., Karrison, T. G., Roegner, R. H. and Williams, D. E. (1990). Health status of Air Force veterans occupationally exposed to herbicides in Vietnam. *Journal of the American Medical Association*, **264**, 1824–1831.

Appendix

We derive the conditional expected value and variance of $\boldsymbol{y}_i$, given that it takes values larger than $\boldsymbol{a}$. We assume that Ω has AR(1) structure. In the following development, for ease of presentation and without loss in generality, we drop the suffix i from $\boldsymbol{y}_i$ and the associated mean $\boldsymbol{\theta}_i$. For convenience we consider the case $\boldsymbol{a}^t = a\,(1,1,\ldots,1)$, the general case follows similarly. It is convenient to derive these results in terms of centered random variable $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{\theta}$. The condition $\boldsymbol{y} > \boldsymbol{a}$ becomes $\boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}$.

Case 1 $k = 3$. We assume that the three dimensional random variable $\boldsymbol{z}$ is normally distributed with mean $\mathbf{0}$ and covariance matrix Ω which satisfies the AR(1) assumption. Then, the conditional distributions of z_1 and z_3 given z_2 are stochastically independent. Consequently, the joint distribution of $\boldsymbol{z}$ can be written as a product of three normal densities,

$$\begin{aligned} f(\boldsymbol{z}) &= (2\pi)^{-\frac{3}{2}} \mid \Omega \mid^{-\frac{1}{2}} \exp[-\frac{1}{2}\boldsymbol{z}^t\,\Omega^{-1}\,\boldsymbol{z}] \\ &= f_1(z_1 \mid z_2) \times f_3(z_3 \mid z_2) \times f_2(z_2), \end{aligned} \tag{1}$$

where

$$\begin{aligned} f_1(z_1 \mid z_2) &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp[-\frac{(z_1-\rho z_2)^2}{2(1-\rho^2)}], \\ f_3(z_3 \mid z_2) &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp[-\frac{(z_3-\rho z_2)^2}{2(1-\rho^2)}], \\ f_2(z_2) &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{z_2^2}{2}) \equiv \phi(z_2), \end{aligned}$$

with $\phi(\cdot)$ representing the density function of the standard normal random variable. This allows us to express

$$\begin{aligned} S(\boldsymbol{a},\boldsymbol{\theta}) &= \int_a^\infty \int_a^\infty \int_a^\infty f(\boldsymbol{y},\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{y} \\ &= \int_{a-\theta_2}^\infty s_1(a,z_2)\ s_3(a,z_2)\ \phi(z_2)\ \mathrm{d}z_2, \end{aligned}$$

where for $j = 1,3$, $s_j(a,z_2) = \int_{a-\theta_j}^\infty f_j(z_j \mid z_2)\ \mathrm{d}z_j = 1 - \Phi[((a-\theta_j) - \rho z_2)/\sqrt{1-\rho^2}]$, where Φ is the standard normal distribution function. Thus, $S(\boldsymbol{a},\boldsymbol{\theta})$ can be evaluated as a one-dimensional integral in z_2. Using (1) we now derive the first two conditional moments.

Conditional expectations: We consider $\mathrm{E}[z_j \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}]$, $j = 1,2,3$. In view of equation (1) these quantities are reducible to one-dimensional integrals as shown below. The evaluation of $\mathrm{E}[z_2 \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}]$ is straightforward,

$$\begin{aligned} \mathrm{E}[z_2 \mid \boldsymbol{y} > \boldsymbol{a} - \boldsymbol{\theta}] &= S^{-1}(\boldsymbol{a},\boldsymbol{\theta}) \int_{a-\theta_1}^{\infty} \int_{a-\theta_2}^{\infty} \int_{a-\theta_3}^{\infty} z_2 f(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} \\ &= S^{-1}(\boldsymbol{a},\boldsymbol{\theta}) \int_{a-\theta_2}^{\infty} s_1(a,z_2) s_3(a,z_2) z_2 f_2(z_2) \mathrm{d}z_2. \end{aligned}$$

It is easy to verify that for $j = 1,3$,

$$\begin{aligned} \int_{a-\theta_j}^{\infty} z_j f(z_j \mid z_2) \mathrm{d}z_j &= [\sqrt{1-\rho^2} h_j(a,z_2) + \rho z_2] s_j(a,z_2) \\ &\equiv \psi_j(a,z_2) s_j(a,z_2), \end{aligned} \tag{2}$$

where

$$h_j(a,z_2) = \phi(\frac{a - \theta_j - \rho z_2}{\sqrt{1-\rho^2}})/s_j(a,z_2).$$

Consequently, for $j = 1,3$,

$$\mathrm{E}[z_j \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}] = S^{-1}(\boldsymbol{a},\boldsymbol{\theta}) \int_{a-\theta_2}^{\infty} s_1(a,z_2) s_3(a,z_2)\, \psi_j(a,z_2)\, f_2(z_2) \mathrm{d}z_2.$$

Conditional covariances: The conditional covariance of $\boldsymbol{z}$ is

$$\mathrm{Cov}[\boldsymbol{z} \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}] = \mathrm{E}[\boldsymbol{z}\boldsymbol{z}^t \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}] - \mathrm{E}[\boldsymbol{z} \mid \boldsymbol{y} > \boldsymbol{a}]\, \mathrm{E}^t[\boldsymbol{z} \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}].$$

The matrix $\mathrm{E}[\boldsymbol{z}\boldsymbol{z}^t \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}]$ has six distinct elements, all of which can be reduced to one-dimensional integrals. Clearly,

$$\begin{aligned} \mathrm{E}[z_2^2 \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}] &= S^{-1}(\boldsymbol{a},\boldsymbol{\theta}) \int_{a-\theta_1}^{\infty} \int_{a-\theta_2}^{\infty} \int_{a-\theta_3}^{\infty} z_2^2 f(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} \\ &= S^{-1}(\boldsymbol{a},\boldsymbol{\theta}) \int_{a-\theta_2}^{\infty} s_1(a,z_2)\, s_3(a,z_2)\, z_2^2 \phi(z_2) \mathrm{d}z_2. \end{aligned}$$

Using (2) it follows that for $j = 1,3$,

$$\mathrm{E}[z_j\, z_2 \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}] = S^{-1}(\boldsymbol{a},\boldsymbol{\theta}) \int_{a}^{\infty} s_1(a,z_2)\;\, s_3(a,z_2)\;\, \psi_j(a,z_2) z_2 f_2(z_2) \mathrm{d}z_2,$$

and

$$\mathrm{E}[z_1\, z_3 \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}] = S^{-1}(\boldsymbol{a},\boldsymbol{\theta}) \int_{a}^{\infty} s_1(a,z_2)\;\, s_3(a,z_2)\, \psi_1(a,z_2) \psi_3(a,z_2) f_2(z_2) \mathrm{d}z_2.$$

To evaluate $\mathrm{E}(z_j^2 \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta})$, we need the following intermediate result. Using integration by parts, we obtain

$$\int_{a-\theta_j}^{\infty} z_j^2 f_j(z_j \mid z_2)\, \mathrm{d}z_j \;=\; t_j(a,z_2) s_j(a,z_2),$$

where

$$t_j(a, z_2) = (1-\rho^2) + \rho^2 z_2^2 + (1-\rho^2)^{1/2} h_j(a, z_2)(a - \theta_j + \rho z_2).$$

Hence,

$$\mathrm{E}[z_j^2 \mid \boldsymbol{z} > \boldsymbol{a} - \boldsymbol{\theta}] \;=\; S^{-1}(\boldsymbol{a}, \boldsymbol{\theta}) \int_{a-\theta_2}^{\infty} s_1(a, z_2)\, s_3(a, z_2)\, t_j(a, z_2)\, \phi(z_2) \mathrm{d}z_2.$$

Case 2 $k = 4$. The joint density of $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{\theta}$ can be written as

$$f(\boldsymbol{z}) = f_{1.23}(z_1 \mid z_2, z_3) f_{4.23}(z_4 \mid z_2, z_3) f_{23}(z_2, z_3),$$

where the f's denote normal density functions. Consequently the first and second moments of the z_i's can be written as two-dimensional integrals.

Case 3 $k \geq 5$. The approach taken above can be applied in this case also. For example, when $k = 5$ it can be verified that the joint density of $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{\theta}$ satisfies

$$f(\boldsymbol{z}) = f_{1.234}(z_1 \mid z_2, z_3, z_4) f_{5.234}(z_5 \mid z_2, z_3, z_4) f_{2.3}(z_2 \mid z_3) f_{4.3}(z_4 \mid z_3) f_3(z_3).$$

Thus the survival function and conditional moments can be reduced to three-dimensional integrals.